NOTES ON

MEDICAL STATISTICS

- Collected By -
Dr. Arunkumar. Biradar

MEDICAL STATISTICS

SYLLABUS POINTS

- Application of statistical methods to Ayurvedic research, Collection, Compilation and tabulation of medical statistics, methods of presentation of data, calculation of mean, Median and Mode of Measurement of variability, Standard deviation, Standard error, Normal probability curve.
- Concept of regression and co-relation and their interpretation.
- Tests of significance, t, x^2, z and f test and their simple application.
- Principle of Medical Experimentation on variations in experimental design.
- Vital Statistics.

GENERAL INTRODUCTION TO STATISTICS

DERIVATION / ORIGIN OF THE WORD STATISTICS

The word statistics is derived from –

- A Latin word – Status.
- A Italian word – Statista.
- A German word – Statistic.

All of these words refer to a political state which is because of reasons that the knowledge of statistics was used to run a State / Kingdom / Country.

According to Webstar –

Statistics is the classified facts representing the condition of the people in a state, specially those facts which can be expressed in terms of numbers / in tables / in a classified.

The word statistics can be used both in singular and plural sense. It gives different understandings when used in singular or plural form.

Singular meaning of Statistics –

Here, it refers to science.

In singular sense, word statistics is used to mean a subject, science or a discipline.

Statistics is a study of knowledge, which deals with different methods of collection, classification, presentation, analysis and interpretation of data.

Data – It refers to the sort of information, which is collected in terms of value.

Plural meaning of Statistics –

According to Secriest, the plural meaning of statistics refers to statistical methods. Viz. –

- Aggregate of facts.
- Affected to a marked extend by multicity of causes.
- Numerically expressed.
- Enumerated / estimated according to reasonable standards of accuracy.
- Collected in a systematic manner.
- For a predetermined purpose / cause.
- Placed in relation with each other.

01. AGGREGATE OF FACTS

It refers to the collection of various data.

e.g. Collection of Blood pressure, weight, height, etc of 20 students in a class.

02. AFFECTED TO A MARKED EXTEND BY MULTICITY OF CAUSES

A sample or a subject or a recording be affected by various internal, external or miscellaneous causes like Age, Sex, Time, Place, Food habits, Religion, etc.

e.g. Blood pressure variation according to the change in emotional status, hormonal changes, etc.

03. NUMERICALLY EXPRESSED

Quantifying the data. (i.e. Expression of the collected data in terms of the values.)

e.g. Blood pressure – 120/80 mm of Hg, 140/90 mm of Hg, etc.

04. STANDARDS OF ACCURACY

Data should be standardized according to the normal values. (i.e. In between the range of minimal and maximal values.)

e.g. Record of blood pressure from 0-300 mm of Hg only.

Variation of +/- 15 mm of Hg in systolic blood pressure.

Variation of +/- 10 mm of Hg in diastolic blood pressure, etc.

05. COLLECTION IN A SYSTEMIC MANNER

For the collection of data various methods of researches should be adopted. (i.e. Standards with a particular restriction)

e.g. Performing dhara only for 40 minutes.

Recording Blood pressure sharply at 09.00 am only.

06. FOR A PREDETERMINED PURPOSE

Collection of data based on research plan / requirement of the researcher. (i.e. according to the aims and objectives of the research project)

e.g. Collection of the blood sugar levels before and after the Madhutailika basti prayoga in 30 diabetic patients.

07. PLACED IN RELATION WITH EACH OTHER

Co-relation of the data collected. (i.e. Co-relation of the data collected before and after the interventions, variables observed during the study like height, place, temperature, etc during the study, etc.)

BRANCHES OF STATISTICS

There are 2 main branches of the statistics –

- ✓ Descriptive statistics.
- ✓ Inferential statistics.

DESCRIPTIVE STATISTICS

It refers to the various statistical measures that are used to describe the various characteristics of data. From this type of statistics we can not conclude over the collected data.

e.g. Mean, Mode, Median, Standard deviation, etc.

INFERENTIAL STATISTICS

It refers to various statistical measures that are used to draw some valid conclusions and findings.

e.g. Test of significance like t-test, f-test, z-test, Chisquare test, etc.

OBJECTIVES OF THE STATISTICS

- ➢ The objectives of statistics are of two folded i.e. To condense, organize and summarize the collected raw data.
- ➢ To reach or draw or to take decisions about a large data (population) by examining a small part (sample) of data.

APPLICATION OF STATISTICS

Science with statistical support will yield fruits. (i.e. will achieve its maximum outcome).

The science of statistics can be applied to any of the scientific fields like economics, politics, industry, business, education, administered medicine and so on.

When the statistical methods or science of statistics are applied for public health, medicine or biological data, it is called as **Medical Statistics** or **Biostatistics** or **Biometry**.

BIOSTATISTICS

Biostatistics, is a subject, which deals with application of statistical methods in the field of medicine, biology and public health in planning or conducting and analyzing data which arise in investigations.

In other words, it is an application of different statistical methods i.e. collection, classification, presentation, analysis, interpretation of biological variations.

It is also known as **Quantitative Science**. Because, in statistics the facts and observations should be expressed in figures or numbers.

The other **synonyms** of Biostatistics is, **Science Of Variation**. Because, it deal with the various dependants and independent variables.

Biostatistics is also known as **Biometry**.

VARIABLE

The characteristics varies in person, time and place is called variable.

As the statistics deals with the variables. So, it is called as **Science of Variables**.

BIOMETRY

It is a Greek word, formed by the combination of 2 words –

Bio + Metry.

Here, Bio is the word related with the Biology or Life.

Metry refers to the Measurement.

So, the word biometry means, the measurement of the life.

Depending upon the application of Biostatistics in various fields it is named as – Health statistics, Medical statistics, Vital statistics, etc.

HEALTH STATISTICS

It deals with the public / community health.

MEDICAL STATISTICS

When the statistics is applied in the field of the medicine, it is called as medical statistics. The action of drugs, various treatment modalities, etc.

VITAL STATISTICS

When the statistics is applied in the field of demography (i.e. Study of the population) and its important events like – Birth, Death, Mortality rate, Fatality rate, etc called as Vital statistics.

APPLICATION OF BIOSTATISTICS IN THE FIELD OF AYURVEDIC RESEARCH

Ayurveda, deals with the four types of Ayu i.e. Hitayu, Sukhayu, Ahitayu, Dukhayu.

हिताहितं सुखंदुःखं आयुःतश्चहिताहितं ।

मानम् च तश्च यत्रोक्तं आयुर्वेद स उच्यते ।। च. सू.

Ayurveda also deals with the measurement.

आयुःकामय मानेन धर्मार्थसुखसाधनम् ।

आयुर्वेद उपदेशेषु विधेय परमादरः ।। **अ. सं. सू. १**

So, it can be concluded that both biometry as well as Ayurveda deals with the measurement of life.

Biometry, can be applied in various fields of Ayurvedic Researches like – Literary study, Pharmacological study, Clinical study, Survey study, etc.

Some of the common applications of the Biostatistics are as follows –

TO SIMPLIFY OR TO CONDENSE THE HUGE DATA

- Collection of the lakshanas of various diseases.
- Collection of lakshanas as per Poorvaroopa, Roopa, Upadrava, Asadhya lakshana, Arishta lakshana, etc. (i.e. Hetu kosha, Lakshana kosha)
- Literary study on Prakriti – Collection of various factors about Prakriti and classifying them according to the physical factors, psychological factors, Shadanga shareera, etc.
- Vyadhi Kshamatwa – Collection of the concept of Bala in various texts and dividing them as per the dividing base i.e. Sahaja bala, Kalaja bala, Yuktikrita bala.

TO TEST THE HYPOTHESIS

Whatever mentioned in classics, to re-evaluate the concept.

e.g. नासा हि शिरसो द्वारम् । तत्रावसेचितं औषधं श्रोत्र श्रृंग्राटकं प्राप्य व्याप्यच मूर्धानं – – – इवासक्ता दोषान् अपहरति ।

Conducting a well planned research work to confirm the above mentioned classical concept through various ways.

Sushruta opines that, the diseases which can be cured by Kavalagraha also cured by Pratisarana. Hence, both the procedures are having equal potency in the treatment of Kanthagata rogas. Conducting a well designed research work to evaluate the same with the same drug with two different procedures can be undertaken.

TO DRAW THE CONCLUSIONS

Based on the conducted or based on previous studies, some conclusions are drawn and if necessary some recommendations are suggested.

e.g. When a scholar planned a research work to evaluate the effect of Kavalagraha in Mukhapaka with some medicine but with varying duration of the Kavalagraha. (i.e. 5 minutes, 10 minutes, 15 minutes, etc.) In this research work finally on the basis of statistical results obtained the scholar can draw some conclusion and can standardize the particular time for the Kavalagraha procedure in respected condition.

TO STUDY THE RELATIONSHIP BETWEEN 2 OR MORE VARIABLES

This can be done with the help of concept of co-relation.

e.g. When a scholar planned a research work to evaluate the effect of Kavalagraha in Mukhapaka with some medicine but with varying duration of the Kavalagraha. (i.e. 5 minutes, 10 minutes, 15 minutes, etc.) In this research work finally on the basis of statistical results obtained the scholar can draw some conclusion and can standardize the particular time for the Kavalagraha procedure in respected condition.

Relation between the age and height.

Relation between the fatty diet and chances of atherosclerosis.

Relation between the number of cigarettes per day and the life span of smokers, etc studies can be undertaken.

TO PREDICT THE FUTURE THINGS (i.e. to assess the future events)

This can be done with the help of the concept of regression.

e.g. Suppose, if we have data of number of cases in Poliomyelitis of last 5 years. Regression analysis can help in prediction of the probable number of cases in the next year.

It is very useful in target setting, Budget sessions, etc.

IN THE FIELD OF VITAL STATISTICS

Vital statistics deals with the important events of life, which are indicative of population or community health.

e.g. It is very important to know about the community health problems and to counter such problems through the various plans and projects.

LIMITATIONS OF STATISTICS

- Statistics deals with the quantitative characters rather than qualitative data.

e.g. Statistics can predict the number of books in library, but not the number of good quality books.

- Statistics does not deal with individual or single character. It is true on average.

e.g. In class A, 3 students scored 35, 35 and 35 marks respectively. The mean score of the class will be 35+35+35=105/3=35.

In class B, 3 students scored 78, 22 and 5 marks respectively. The mean score of the class will be 78+22+05=105/3=35.

Though, the average is same in both the groups, the individual values differs. This is the limitation of the statistics. Here, statistics deals with the group not with an individual entity. Though the average marks scored in both classes is same it does not mean that all the students have scored similar marks. But, this limitation can be neglected / nullified by the concept of dispersion.

- Statistical results may be hampered by various physical, biochemical, analytical, methodology, etc. forms of research bias. (i.e. Errors in conducting research.)

e.g. Errors done by researchers, Errors in methodology, Errors in analysis, Errors in collection and calculation of data, etc.

- Statistics can be miss used and wrong statistical methods can be manipulated.

e.g. “Number of accidents are committed by females are less as compared to Males.” Out of 1000 male riders, 15 males were committed with accident. Out of 100 female riders, 3 were committed with accident. Here, numerically the number of accident seems to be more in males, but it is wrong to give above mentioned statement. Because, the incidence of the event taken in both the group is not same. If we take the mean in male riders it will be 1.5 and in females it will be 3.0. So, if we calculate the incidence as per the size of population the number of accidents committed by females will be 30. It is clear that, female riders are more prone to commit accidents. So, the above mentioned statement is statistically wrong.

COLLECTION OF DATA

DATA

It refers to the given piece of information. In other words, it is aggregate of figures, numbers or the set of the values i.e. recorded in one or more observational queries.

OBSERVATIONAL UNITES

The source of observation is called as observational unites.

e.g. Such as object, person, patient, etc.

OBSERVATIONS

The combination of events and its measurement constitute observation.

e.g. Measuring the Blood pressure is the event & the measured blood pressure like 102/80 mm of Hg will be measurement. The combination of both event and measurement i.e. Observation.

Features / Characteristics of an Ideal Data

It should be – **(CURA)^2**

- Complete
- Comparable.
- Up to dated.
- Understandable.
- Reliable.
- Relevant.
- Accurate.
- Available easily.

CLASSIFICATION OF DATA

Data is classified on various basis as mentioned below –

Based on the characters ⟹ Qualitative.
⟹ Quantitative.

Based on Method of collection ⟹ Continuous.
⟹ Discrete.

Based on Classification ⟹ Primary.
⟹ Secondary.

CLASSIFICATION OF DATA BASED ON THE CHARACTERS

QUALITATIVE DATA

It is also called as Attribute / Character.

It is a data, where character or quality is constant, but frequency varies.

This is always represented in the form of discrete or discontinued and countable.

e.g. Sex, Religion, Nationality, etc.

In a class number of students is fixed. Classification of students on the basis of sex, which is a fixed character, and it is countable called as qualitative data.

Out of 20 students, 21 are male and 08 are female students. Here, total number of male can not be 18.2, 18.5 like that total number of female can not be 08.6, 08.9.

QUANTITATIVE DATA

In this type / set of data character as well as frequency varies.

e.g. Following are the heights of people aging between 10 to 20 years.

Sl.	Height (In feats)	Frequency
01.	**3 – 4**	10
02.	**4 – 5**	20
03.	**5 – 6**	10

Here, both frequency and character changes. Out of 40 people height frequency is mentioned above. 20 people found in 4 – 5 feats character. It means, 20 people height lies between 4 – 5 feats. Then it may be 4.1, 4.2, 4.3, etc.

This type of data called as Discrete and continuous in nature.

CLASSIFICATION OF DATA BASED ON METHOD OF COLLECTION

DISCRETE DATA

The data collected by the method of counting and representing in round numbers and integral, is called as discrete data.

e.g. Number of patients visiting O.P.D.

Sl.	Day	Number of Patients
01.	Monday	210
02.	Tuesday	250
03.	Wednesday	450

Here, the number of patients can not be 210 ½, 210 ¾ like that. So, this type of countable data called as discrete data.

CONTINUOUS DATA

The data which is collected by using measuring instrument and represented as round number or fraction or decimals, is called as continuous data.

e.g. Weight of New borns in a hospital – 2.8 Kg, 3.5 kg, etc.

Hb% of the patients – 8.6gm%, 11.5gm%, etc.

CLASSIFICATION OF DATA BASED ON FUNCTIONAL CLASSIFICATION

PRIMARY DATA

Those data, which are collected for the very first time, original in nature under the control and supervision of medical investigator, is called as primary data.

e.g. A research scholar collecting data for thesis work. Number of family planning operations conducting in P.H.C., etc.

SECONDARY DATA

The data which is not collected by the investigator, but it is derived from other reliable sources, referred as secondary data.

e.g. The D. H. O. collects the information about the number of Tuberculosis patients in a district.

A doctor wants to study the relationship of smoking and Heart diseases based on the data given in Indian Medical Journals, etc.

RELIABLE SOURCE OF DATA

The data which is collected from a reliable source like Government offices, Standard and Recognized institutes, National and International Organization, etc.

The National Level – Various ministries coming under Government of India.

e.g. Ministry of Family and Health Welfare, Ministry of Mother and child Health welfare, etc.

The State Level – Various ministries running under the state Government under the control of Central Government.

The District Level – District / Community hospitals running under the control of state government respective ministries.

The Local Level – Recognized hospitals, NGO's, Private organizations, etc

The various standard Index Journals and Publications like BMJ, etc.

VARIABLE

A characteristic that takes on different values in different persons, places or things.

CONSTANT

Quantity that do not vary in a given set of observational data. they do not require statistical study. (S.D., S.E., Mean, C.C.)

POPULATION

Study of elements such as person, things or measurements for which we have an interest at a particular time.

SAMPLE

Part of population or group of sample unit.

SAMPLING UNIT

Each member of a population.

PARAMETER

Summary value or constant of a variable that describe the population such as mean, C. C., etc.

STATISTIC

Summary value that describe the sample such as its mean, S.D., S.E., etc.

PARAMETRIC TEST

It is one in which population constants are used such as mean, variance, C.C., etc.

NON-PARAMETRIC TEST

The tests such as x2 test in which population no constant of a population is used. Data do not follow any specific distribution and no assumptions are made. **e.g.** To clarify good, better, best values.

COLLECTION OF DATA

DEFINITION

The various methods by which the necessary samples or data are collected for the study in a systemic manner depending upon need / requirement of researcher.

SOURCE OF COLLECTION OF DATA

There are main 3 sources.

- Experiments
- Surveys
- Records

EXPERIMENTS

Various experiments are conducted for investigation and fundamental research based on the basic principles of particular science.

The data is collected with specific objectives and the results obtained are used in the preparation of dissertation, thesis, research paper, journal articles, etc.

SURVEY

It is used in epidemiological studies to find out the incidence or prevalence of health or disease in a community.

Survey provide useful information for –

- Changing the trends in health status, morbidity, mortality, etc.
- Provides feed back, which will be helpful to plan or alter or to modify the policies run by Government or any of the authority.

RECORDS

These are maintained for a long period of time in registers or books of concern departments like Central Government, State Government, etc.

These are used for various purposes like Vital statistics, demography, etc.

METHODS OF COLLECTION OF DATA

It is important to differentiate a primary or a secondary data before we start the collection. The important methods of collection of data are –

- Observational
- Interview
- Questionnaire
- Experimental

OBSERVATIONAL METHOD OF DATA COLLECT

The general observation does not stand for observation.

Observation is a scientific toll and a systematic method of collection of data (i.e. In preview of the objective of the researcher.)

Types

Based on systematic plan and organization of the researcher, the observation is divided into 3 categories –

- Structured
- Unstructured

STRUCTURED OBSERVATION

If the data collection is done in a systematic manner, with fulfillment of all pre-requisites, then it is called as Structured Observation.

Most of the researches use this type of observation.

UNSTRUCTURED OBSERVATION

If a systematic approach is not taken towards data collection, it is called as unstructured observation.

Types of Observation

Based on the involvement of observer, observation it is divided into –

- Participant Observation
- Non-participant Observation

PARTICIPANT OBSERVATION

When the observer becomes a part of the sample, understanding in the emotional, socio-cultural, occupational background, it is called as Participant Observations.

e.g. A research scholar conducting a research in his native area, called as Participant observation. Because, the observer will be the native of that particular area and will be aware with all the emotional, socio-cultural, occupational background of the samples.

NON PARTICIPANT OBSERVATION

When the observer is not a part of the sample and there will not be any understanding in the emotional, socio-cultural, occupational background, it is called as Non-participant Observations.

In this type of observation, the chances of bias is more.

e.g. A Indian research scholar conducting a research in London which is totally different from his present status, called as Participant observation. Because, the observer will not be the part of that particular area and will not be aware with all the emotional, socio-cultural, occupational background of the samples.

Benefits / Merits

- Subjective bias is eliminated in participant.
- Independent of willingness by respondent.
- Non-need of active co-operation.

De-merits

- Limited information.
- Same unforeseen factors / Hidden factor may interfere with observation.

INTERVIEW METHOD

It is a form of interrogation / communication based on stimuli and response or questions and answers.

It is of 2 types –

- Direct personal investigation.
- Indirect oral examination.

DIRECT PERSONAL INVESTIGATION

It is a form of investigation where the interviewer relies on the wordings of the interviewee.

INDIRECT ORAL EXAMINATION

It is a form of examination, where the cross check of the interview is done by related person.

e.g. Paediatric examination, Psychiatric examination, CBI investigations, etc.

Characteristics of Interviewer

Interviewer should be – Polite, honest, sincere, impartial, technical, competence with necessary practical experience and must be friendly with the interviewee.

Guidelines for interviewer

- Interviewer should know the problem and well planned prepared.
- Always have good set up. (Cool and Calm)
- Have friendly and informal talks.
- Have curiosity and respect.
- Ask well phrased questions.
- Should not hurt the interviewee.
- The matter must be confidential.

Merits

- More detail information can be obtained.
- Greater flexibility to restructure the questions.

De-merits

- Respondent / Subjective bias.
- Time consuming.

QUESTIONNAIRE METHOD

It is a method, where the questions are given and the respondent is asked to reply the same according to the instructions.

It is of 2 types –

- Given
- Posted

GIVEN

In this type of questionnaire method a set of questions is prepared and provided to the respondent. Sufficient time is given to respondent to answer the given questions.

POSTED

In this type of questionnaire method a set of questions are prepared and provided to the distant respondent. Sufficient time is given to respondent to answer the given questions and asked the respondent to post it back to the observer. In this type of method there is low return rate.

GUIDELINES FOR QUESTIONNAIRES

- Questions should be simple, clear, understandable and related to the topic or problem.
- Decide either closed end or open end or even both types of questions.
- Maintain the sequence (order) of questions (i.e. From general to complex)
- Questions should not be related to personal character / wealth.
- Questions should not hurt the person.
- Avoid the use of those questions which puts too much of strain to one's memory or intellect. (i.e. it should be according to the qualification and I. Q. of the respondent.

Merit

- Time saving.
- Low cost.
- Large sample can be taken.
- Sufficient time to answer.
- Best method to those who are not approaching.

De-Merits

- Can be used in only educated and co-operative patients.
- Low return rate, especially in posting method.
- Doubt about its own version.

EXPERIMENTAL METHOD

The method in which various experiments or measurable instruments are adopted for the collection of data, is called as Experimental method.

Merits

- An ideal objective parameter.
- Beneficial in comparison.
- Lack of subjective bias.

De-merits

- Expensive.
- Chance of observer bias.
- Sometimes it may false positive results.

Hence, it is very important to co-relate the investigative values with the clinical presentations.

COMPILATION & TABULATION OF MEDICAL STATISTICS & METHODS OF PRESENTATION OF DATA

It includes sorting (i.e. classification and presentation of data.)

CLASSIFICATION

Definition

The grouping or arranging or division of data based on some similar or dissimilar characteristics, to facilitate easy analysis and condensation of huge data is called as classification of data.

Types

Based on the number of attributes / characteristics it is divided into 2 types.

- Simple
- Manifold

SIMPLE CLASSIFICATION

If the classification is based on the single attribute / characteristic is called as simple classification.

e.g. Single classification based on any of the based entity Age, Sex, Religion, Nutritional status, etc.

Table showing the number of patients in different age groups.

Sl.	Age groups	Number of patients
01.	**10-20**	15
02.	**20-30**	23
03.	**30-40**	24

MANIFOLD CLASSIFICATION

If the classification is based on the 2 or more than 2 attributes, it is called as Manifold classification.

e.g. Single classification based on Age, Sex, Religion, Nutritional status, etc.

Table showing the number of patients according to sex, age groups and their nutritional status.

Sl.	Sex	No. of Pt.'s	Age	No. of Pt.'s	Nutritional status	No. of Pt.'s
01.	Male	**30**	Children	**26**	Normal nutrition	**08**
					Under nutrition	**16**
					Over nutrition	**02**
			Adulthood	**36**	Normal nutrition	**19**
					Under nutrition	**12**
					Over nutrition	**05**
			Adult	**48**	Normal nutrition	**32**
					Under nutrition	**15**
					Over nutrition	**01**
02.	Female		Children	**26**	Normal nutrition	**19**
					Under nutrition	**12**
					Over nutrition	**05**
			Adulthood	**36**	Normal nutrition	**32**
					Under nutrition	**15**
					Over nutrition	**01**
			Adult	**48**	Normal nutrition	**08**
					Under nutrition	**16**
					Over nutrition	**02**

There are 4 important basis of classification of data. viz.

- Quantitative
- Qualitative
- Geographical
- Chronological

QUANTITATIVE DATA

The classification based on numbers or figures, called as Quantitative data.

e.g. Height, Weight, Hb%, Blood pressure, etc.

QUALITATIVE DATA

The classification of data based on the attribute or character, called as qualitative data.

e.g. Sex, Religion, Nationality, etc.

GEOGRAPHICAL DATA

The classification of data is based on the area or place, called as Geographical data.

e.g. Continent, Country, State, District, Takula, Village, etc. Number of tuberculosis patient in each state of India.

CHRONOLOGICAL DATA

The classification of data is based on the duration or time, called as Chronological data.

e.g. Classification of data based on minutes, hours, days, weeks, months, years. etc. Duration / Chronicity of RA in years / months.

OBJECTIVES / USES OF CLASSIFICATION

- To condense the huge data.
- Useful in comparison.
- Simple and easy to understand.
- It refers to systematic representation.
- Can be used for further statistical applications like presentation and analysis of data collected during any research work.

PRESENTATION OF DATA

Definition

Systematic representation of the data, which is collected and classified in the form of tables or drawing (graphs / diagrams) is called as presentation of data.

IDEAL PRESENTATION

- It should be simple and systematic to arouse the interest.
- It should be concised, but there should not be any vomition / deletion of data.
- It should be arranged in logical or chronological manner.
- It should be useful for further analysis.

OBJECTIVES / USE OF PRESENTATION OF DATA

- Easy and better understanding.
- Helpful in future analysis.
- Easy for comparison.
- It gives a first hand information.
- It is an attractive and appealing way of presentation.

Types of presentation

Presentation can be made in mainly 2 forms –

- Tables (Tabulation / Frequency Distribution Tables. FDT)
- Drawing (Geographical Presentation / Frequency Distribution Drawing. FDD)

TABULATION / FREQUENCY DISTRIBUTION TABLE / FDT / TABLES

The systematic presentation of data in rows and columns, called as FDT (Frequency Distribution Table / Tabulation)

Tabulation is a process by which a data of a long series of observation are systematically organized and recorded, so as to unable analysis and interpretation.

CHARACTERISTICS OF FREQUENCY DISTRIBUTION TABLE (FDT)

- It should be simple and clear cut.
- The title of the Frequency Distribution Table (FDT) should be expressed in appropriate terms.
- The figures / numbers in the body of table should be arranged in logical manner.
- If several points are emphasized from the same data, make many small tables.

TYPES OF FREQUENCY DISTRIBUTION TABLE (FDT)

Depending upon the data

It is of 2 types –

- Discrete Frequency Distribution Table (FDT)
- Continuous Frequency Distribution Table (FDT)

DISCRETE FREQUENCY DISTRIBUTION TABLE (FDT)

The table which represents the discrete qualitative or countable data called as discrete Frequency Distribution Table (FDT).

GUIDELINES FOR THE CONSTRUCTION OF DISCRETE FREQUENCY DISTRIBUTION TABLE (FDT)

- Pick the lowest and highest observations.
- Arrange in logical order. (Preferably in ascending order i.e. 0 – 1 – 2, etc.)
- Mark the tally marks against the observations.
- Count the tally marks and write it in frequency / countable data.

e.g. Number of children per family of 15 couples.

Sl.	Observation (x)	Tally marks	Frequency (f)
01.	0	||	2
02.	1	||||	4
03.	2	𝍸	6
04.	3	||	2
05.	1	|	1

In the above mentioned table the number of children is countable. There will not be any family with some 2.5, 5.6 number of children. Such type of presentation of data is called discrete Frequency Distribution Table (FDT).

CONTINUOUS FREQUENCY DISTRIBUTION TABLE (FDT)

The Frequency Distribution Table (FDT) represents the continuous quantitative or measurable data, called as Continuous Frequency Distribution Table (FDT).

e.g. Table showing the marks scored by 15 students.

Sl.	Observation (x)	Tally marks	Frequency (f)
01.	10-20	||	2
02.	10-20	||||	4
03.	20-30	𝍸	6
04.	30-40	||	2
05.	40-50	|	1

In the above mentioned table the number of marks is arranged in groups. There will be varying number of students in each group and the students in a group will not be having same scoring of marks. The number of marks will be in limit the particular class width and the marks can be fractions. Such type of presentation of data is called continuous type of Frequency Distribution Table (FDT).

Guidelines for constructing continuous Frequency Distribution Table (FDT)

- Select the lowest and highest observation.
- Select the suitable width. (i.e. Class width & Class interval)
- Divide the observations into sufficient number of classes. (Preferably in between 5 to 15 classes)
- Make / Mark tally marks (to minimize the mistakes during counting and classifying the huge data in particular groups) and write the frequency against each class.

  Continuous frequency distribution table consists of following entities –
  - Class
  - Class interval
  - Lower limit
  - Upper limit
  - Class mid point
  - Class frequency

CLASS

It is a quantitative classification of data in groups, when the samples are large in number.

e.g. 0-10, 10-20, 20-30, 40-50, etc.

CLASS INTERVAL

It represents the width or the size of the class. It can be calculated by 3 methods –

- Upper limit of the class – Lower limit of the same class.
- Lower limit of the class – Lower limit of the previous class.
- Upper limit of the class – Upper limit of the previous class.

It is always better to calculate the class interval by lower limit of the class from lower limit of the previous class. Because, calculation of the class interval by first method gives false answer in case of inclusive type of table.

e.g. In the class 0-10 and 10-20 the class interval can be calculated by 3 methods.

- Upper limit of the class – Lower limit of the same class. (10 – 0).
- Lower limit of the class – Lower limit of the previous class. (0 – 10).
- Upper limit of the class – Upper limit of the previous class. (10 – 20).

LOWER LIMITS

It is a starting / first value of the class.

e.g. In the class 20-30, 20 is the lower limit of the particular class.

UPPER LIMIT

It is a last / ending limit of the class.

e.g. In the class 20-30, 30 is the upper limit of the particular class.

CLASS MID POINT

It is a single representative value of the class, which is used for the further statistical classification.

It is calculated by 2 methods.

$$\frac{\text{Lower limit + Upper limit}}{2} \qquad \frac{\text{Lower limit (of 1}^{st}\text{ Class) + Lower limit (of next class)}}{2}$$

In the class 20-30, the class mid point will be –

20+30 = 50/2 = 25.

In the class 20-30, 30-40 the class mid point will be –

20+30 = 50/2 = 25.

Among these 2nd method of calculating the class mid point is the better way for inclusive type of tables.

CLASS FREQUENCY

The number of observation following in a particular class called as class frequency.

The sum of all class frequencies will give the total number of observations.

Class frequency of 20-30 is 6.

METHOD OF CONSTRUCTION OF CLASSES

There are 3 methods in constructing classes.

- Exclusive
- Inclusive
- Open end method

EXCLUSIVE METHOD

Upper limit of the class is excluded. (i.e. Not a part of from particular class.) The upper limit of the class will be the lower limit of the next class.

It is used for discrete or continuous type of data.

e.g. 0-10, 10-20, 20-30, etc. Here, there is continuation of the upper limit of one class with the lower limit of the next class.

INCLUSIVE METHOD

The upper limit of the class is included. (i.e. It is a part of the same class.)

Upper limit of the class will not be the lower limit of the next class. Because, it is included in the same class itself.

It is used for discrete data.

e.g. Weight, Hb%, height of the person.

OPEN END

When the lower limit of the first class or upper limit of the last class or both will not be fixed, called as open end method.

It is used to accumulate a few extreme low or high.

e.g. 0, 3, 5, 50, 20, 27, 26, 244487, 6, 89, 984526.

TYPES OF TABLES / FREQUENCY DISTRIBUTION TABLE

There are 3 common types of frequency distribution table (FDT).

- Ordinary frequency distribution table (FDT)
- Relative frequency distribution table (FDT)
- Cumulative frequency distribution table (FDT)

ORDINARY FREQUENCY DISTRIBUTION TABLE (FDT)

It is a type of frequency distribution table (FDT) in which the observations / classes are arranged with their respective frequencies, called as ordinary frequency distribution table (FDT).

Uses :

It is simple, easy understanding for a large data in a snap.

RELATIVE FREQUENCY DISTRIBUTION TABLE (FDT)

It is a type of frequency distribution table (FDT) in which the frequency of each is expressed in terms of fractions, decimals or percentage, is called as relative frequency distribution table (FDT).

It is calculated by the number of frequency of the class divided by the total number of frequencies.

Uses :

It facilitates the comparison of 2 or more sets of data.

It constitutes the basis of understanding the concept of probability.

CUMULATIVE FREQUENCY DISTRIBUTION TABLE (FDT)

It adds the frequency starting from the first class to the last class.

The cumulative frequency of the given class represents the total of all previous class frequency including that particular class.

Uses

To calculate more than and less than values of a given observation / class.

For further statistical calculations like median.

e.g. Table showing the marks scored by 20 students.

Sl.	OFDT (f)	RFD	%	CFD
01.	2	2/20=0.1	10	02
02.	3	3/20=0.15	15	05
03.	2	2/20=0.1	10	07
04.	10	10/20=0.5	50	17
05.	3	3/20=0.15	15	20
5	**20**	**1.0**	**100**	**20**

PROBLEM

An administrator of a hospital has recorded the amount of time a patient waits before being treated by the doctor in O.P.D. The waiting time in minutes are – 12, 16, 21, 20, 24, 3, 15, 17, 29, 18, 20, 4, 7, 14, 25, 1, 27, 15, 16, 5. (= 20 patients). Prepare the various forms of continuous frequency distribution tables.

Answer :

Step 1 : Select the lowest and highest values.

Lowest value among the raw data is 1 and highest value among the raw data is 29.

Step 2 : Prepare the classes.

Total duration lies in between the 1 to 30 minutes.

To prepare 5 classes – 30/5=6.

So, the class interval should be of 6. So, the classes will be 1-6, 6-12, etc.

Step 3 : Preparation of the table.

Title : The Table showing amount of time a patient waits before being treated by doctor in O.P.D.

Sl.	Class	Tally marks	OFDT (f)	RFD	%	CFD
01.	01-06	||||	4	4/20=0.2	20	04
02.	06-12	|	1	1/20=0.1	10	05
03.	12-18	𝍸 ||	7	7/20=0.3	30	12
04.	18-24	||||	4	4/20=0.2	20	16
05.	24-30	||||	4	4/20=0.5	20	20
5	**5**		**20**	**1.0**	**100**	**20**

FREQUENCY DISTRIBUTION DRAWINGS

Presentation of the data in a form of graph or diagram is known as drawing or Geographical presentation or Frequency Distribution Diagram.

Generally, graphs are used to represent quantitative data, where as diagrams are used to represent qualitative data.

GRAPH

These are commonly used frequency distribution drawings. These are of 6 types. Viz. –

- ✯ Histogram
- ✯ Frequency polygon
- ✯ Frequency curve
- ✯ Line graph (Chart)
- ✯ Cumulative frequency diagram (Ogive)
- ✯ Dot or scattered diagram

HISTOGRAM

It is also called as Block Diagram. It is a type of Area diagram where the variable or characters are plotted in X axis (Abscissa) where as frequencies are marked in Y axis (ordinate).

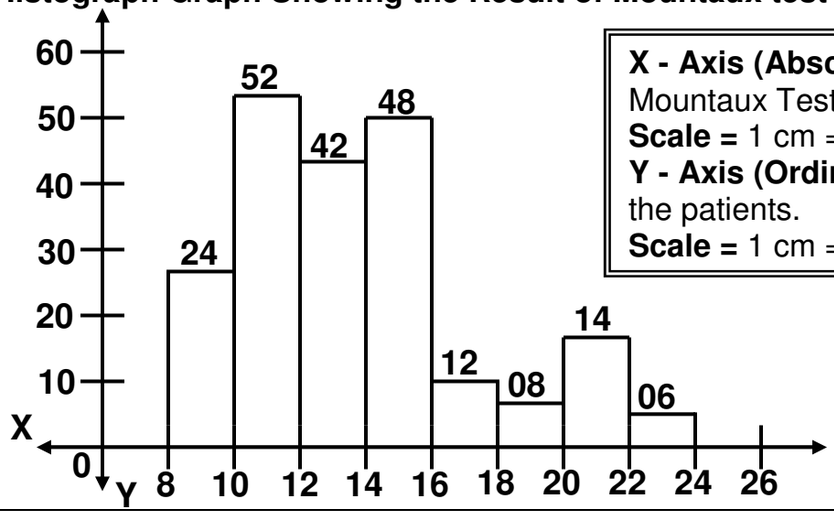
A continuous series of rectangles are formed and this is called as Histogram. The width of the bars may vary.

e.g. Mountaux test of 206 patients.

Result of Montaux test in 206 patients is as follows -

Result of the Test	Number of patients	Result of the Test	Number of patients
08 – 10	24	16 – 18	12
10 – 12	52	18 – 20	8
12 – 14	42	20 – 22	14
14 – 16	48	22 – 24	6

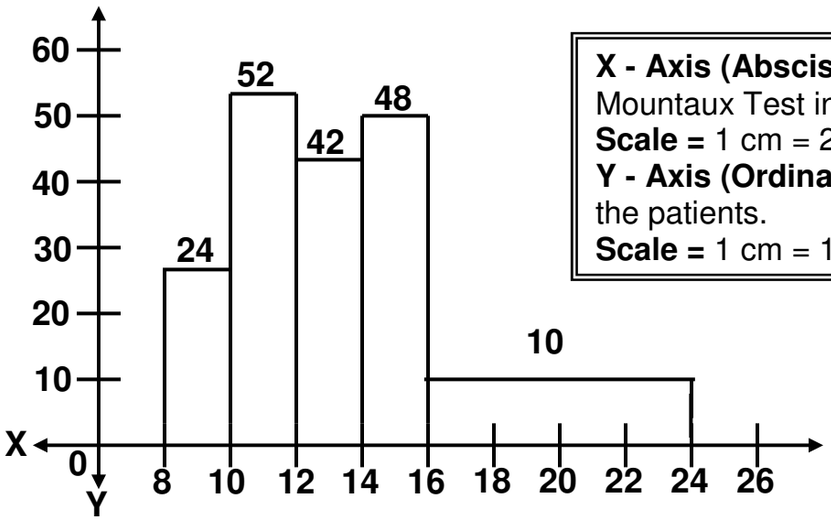
Histograph Graph Showing the Result of Mountaux test in 206 patients.



X - Axis (Abscissa) = Result of Mountaux Test in mm.
Scale = 1 cm = 2 mm.
Y - Axis (Ordinate) = Number of the patients.
Scale = 1 cm = 10 patients.

If we club the groups or classes from 16 - 24 mm in the above group, then the width of the Histogram will vary. Representation of frequency will be done by adding the frequencies of clubbed groups divided by number of classes.

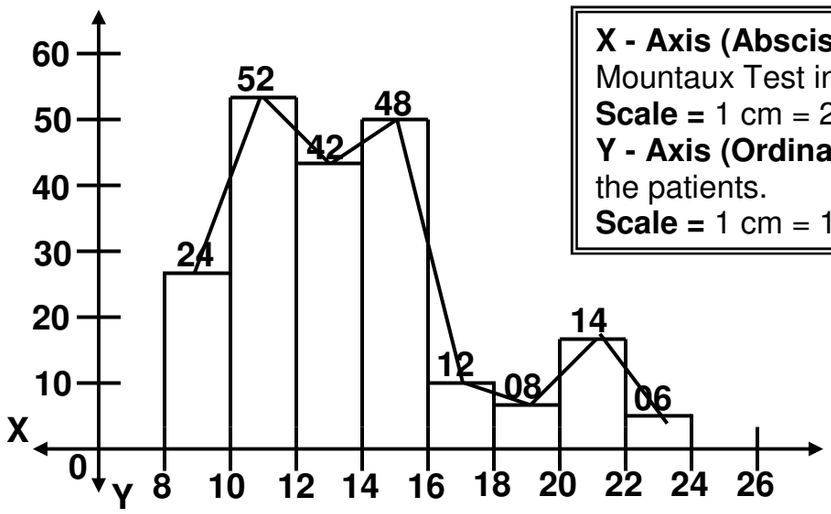
Histograph Graph Showing the Result of Mountaux test in 206 patients.



FREQUENCY POLYGON

Polygon means figures with the many angles. Joining the midpoints of class intervals at the height of frequency after Histogram with a straight line is called as frequency polygon.

Histograph Graph Showing the Result of Mountaux test in 206 patients.



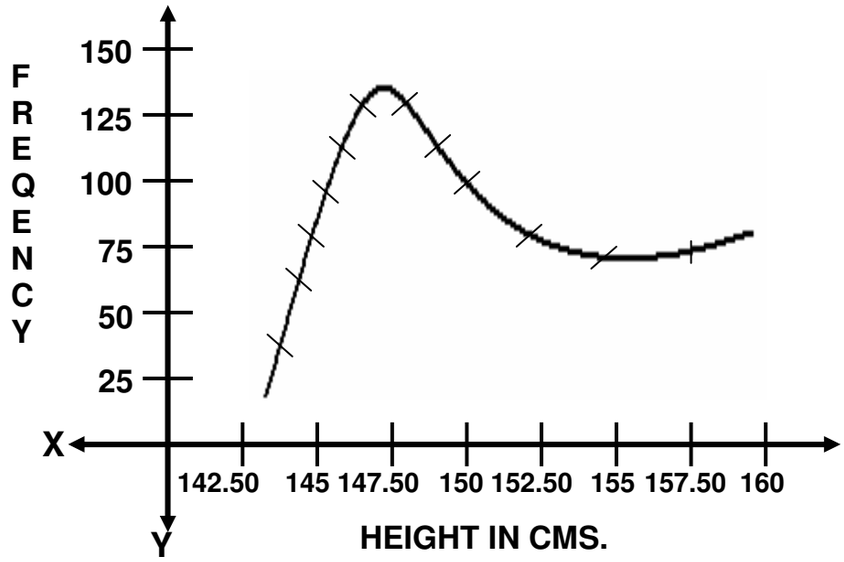
FREQUENCY CURVE

Joining the midpoint of class of frequency without histogram with a smooth curve is called as frequency curve.

Frequency Curve = Frequency Polygon – Histogram.

It is used when there are large numbers of observations.

Frequency Curve showing the Mountaux test result in 206 patients.

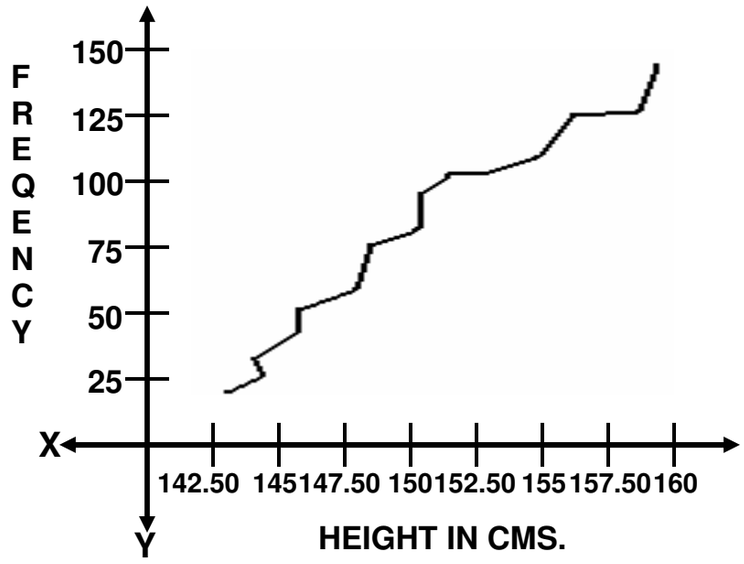


LINE GRAPH OR CHART

The points are marked corresponding to each class or variables against their frequencies and they are joined by smooth line.

It is used to represent the trend in the form of increase or decrease or the fluctuation of given data.

e.g. Population in million of various decades. (It can be either in descending or ascending)



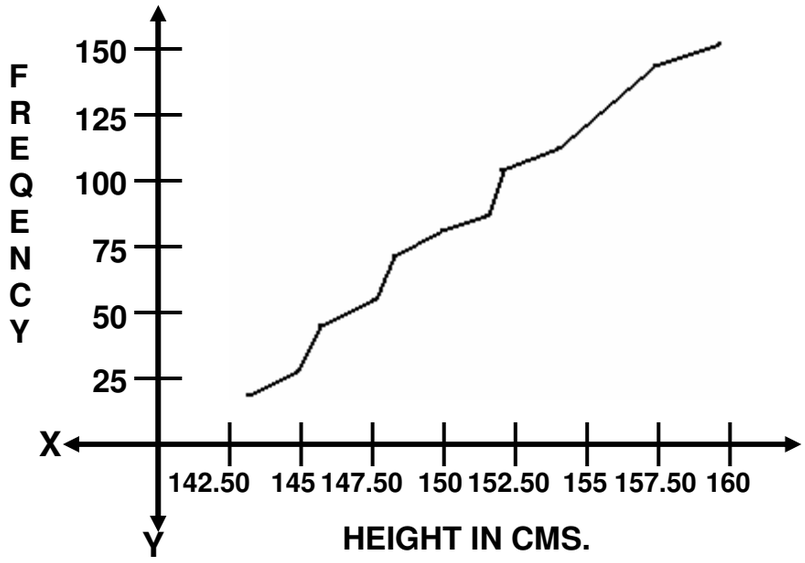
CUMULATIVE FREQUENCY DIAGRAM (OGIVE)

Cumulative frequency diagram is based on cumulative and relative frequency distribution. Before drawing Ogive one has to construct a cumulative frequency distribution table. Later on the diagram is constructed based on variable and its corresponding cumulative frequency. The diagram is drawn bby joining these points with a smooth curve is called as Ogive.

It is used to represent the various percentile like decile (10), quartile (40), pentalile (50), etc.

e.g. Following are the heights of students in a colony. Plot a cumulative frequency diagram for the following data.

SL.	CLASS (HEIGHT IN CMS)	FREQUENCY	CUMULATIVE FD
01.	140 – 145	100	10
02.	145 – 150	150	25
03.	150 – 155	75	42
04.	155 – 160	20	61



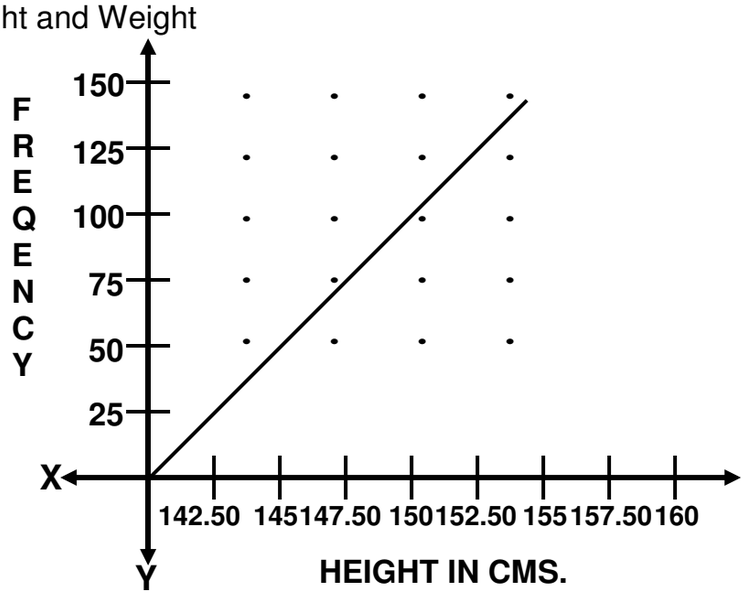
DOT DIAGRAM / SCATTERED DIAGRAM

Generally used in correlation when there is more than one variable to compare this type of diagrams are used.

It is applicable when one has to represent two variables in same direction. One variable can be represented in X axis and other can be in Y axis. We plot variables in X axis, then frequency to be considered in Y axis and viceversa.

It is used in context of correlation. Therefore, it is also called as "Correlation Diagram."

e.g. Height and Weight



PRESENTATION OF QUALITATIVE DATA

To present qualitative or discrete data diagrams are generally used. The commonly used diagrams are as follows –

01. Bar Diagram
02. Pie Diagram – Sector Diagram
03. Pictogram – Picture Diagram
04. Map Diagram – Spot Map

BAR DIAGRAM

Representation in the form of rectangles with spacing with uniform width of rectangle is called as **Bar Diagram**. The spacing between the two bars should be ½ of the width of the rectangle.

Types of Bar Diagram

01. Vertical Bar Diagram
02. Horizontal Bar Diagram

In case of horizontal bar diagram, variable is represented in Y axis and in case of vertical bar diagram variable is in X axis and frequency in Y axis.

e.g. Attendance of Boys and Girls of 1st year PG class.

Bar diagram can be also classified as –

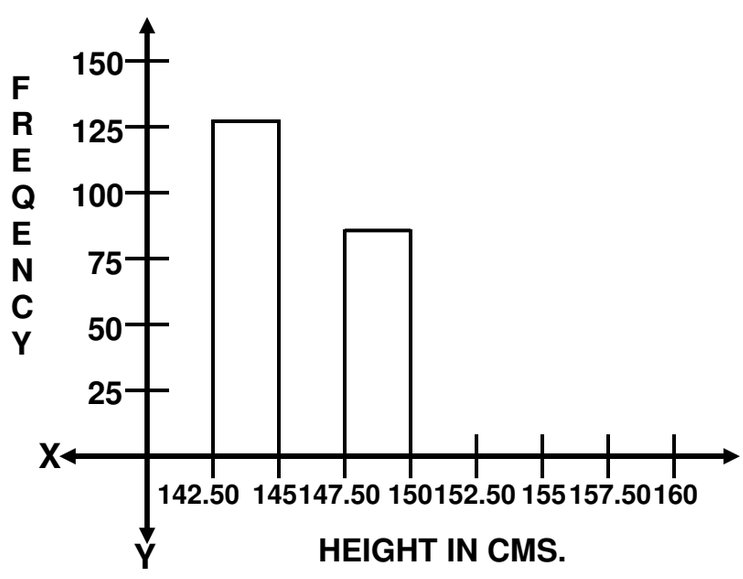
01. Simple bar diagram
02. Multiple bar diagram
03. Proportionate bar diagram

SIMPLE BAR DIAGRAM

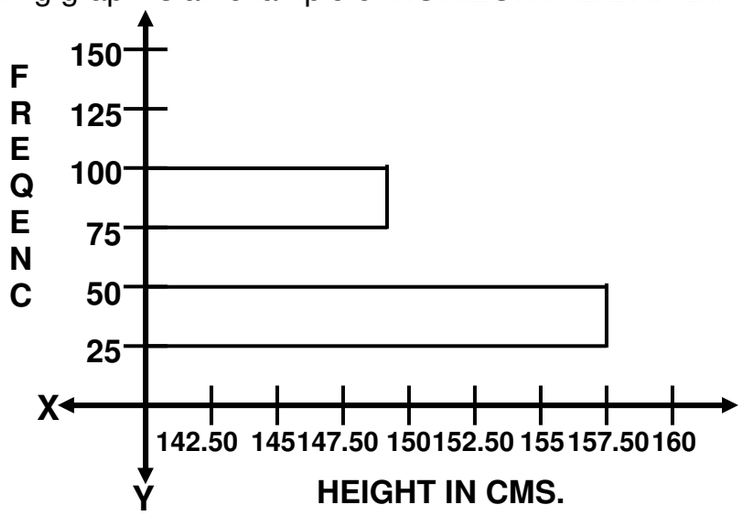
When you represent a single variable as a set of rectangle is called as simple bar diagram.

e.g. Height of Boys of 1st year PG class.

The following graph is an example of **VERTICAL BAR DIAGRAM.**



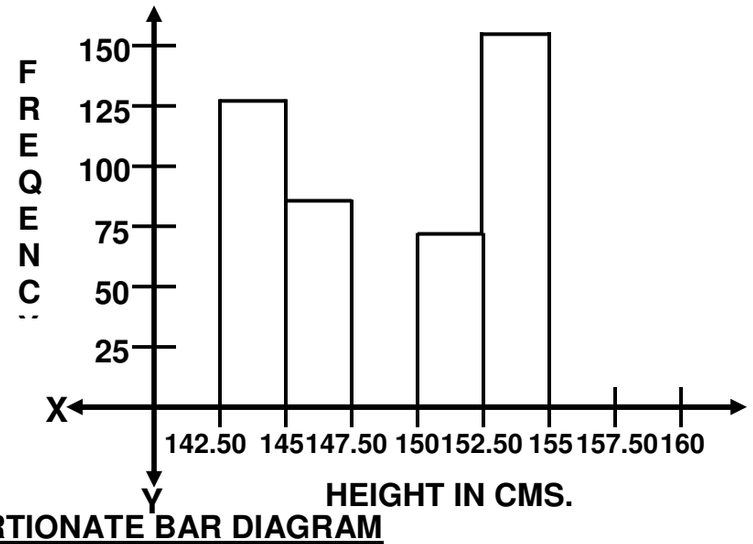
The following graph is an example of **HORIZONTAL BAR DIAGRAM.**



MULTIPLE BAR DIAGRAM

When variables are represented in sets of more than one is called as multiple bar diagram.

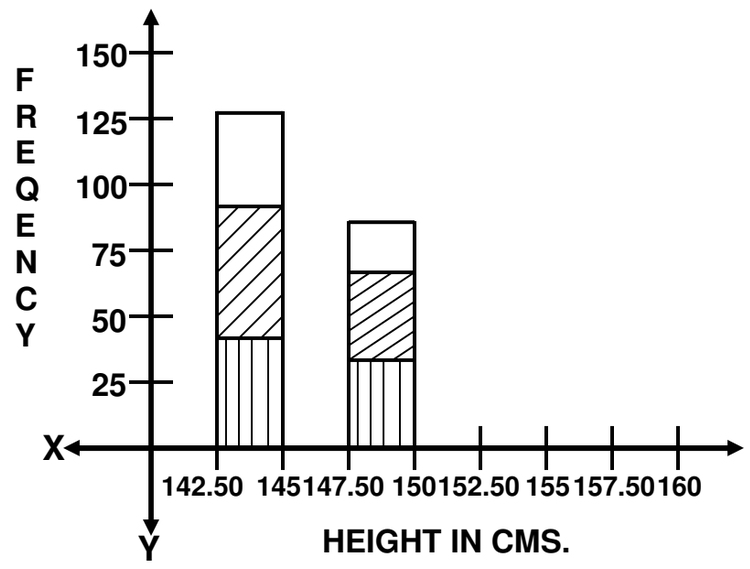
e.g. Heights of boys in 1st, 2nd year PG.



PROPORTIONATE BAR DIAGRAM

Useful for comparison and is represented by subdivision in a same rectangle.

e.g. Heights of boys in 1st,2nd and 3rd year PG classes.



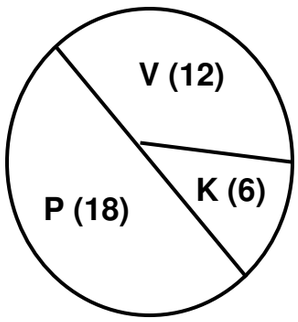
PIE DIAGRAM

It is also called as sector diagram. Frequencies are represented by a circle where each class or observation is represented by class frequency divided by total number of observations and multiplied by 360.

$$\text{Pie Diagram} = \frac{\text{Class frequency}}{\text{Total number of observation}} \times 360$$

e.g. Draw a pie diagram of following data.

Prakriti	Frequency	Calculation	Degrees
Vata	12	12 / 36 x 360	120
Pitta	18	18 / 36 x 360	180
Kapha	6	6 / 36 x 360	60



PICTOGRAM (PICTURE DIAGRAM)

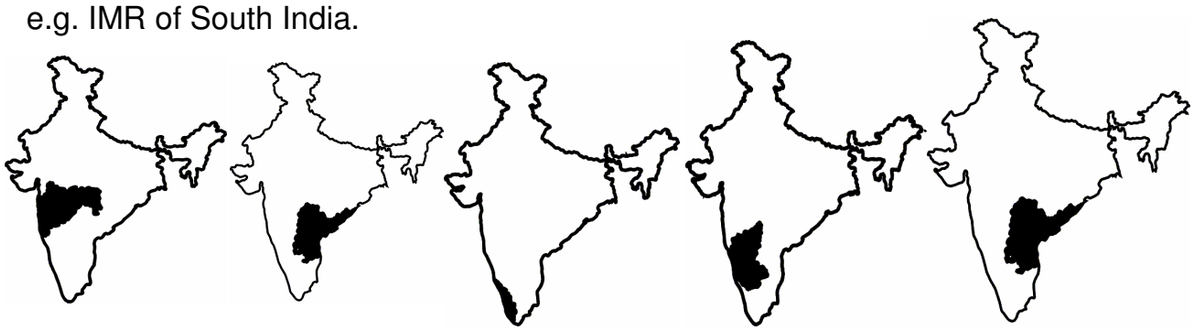
Most common diagram to impress the population. In this diagram actual pictures are used to represent the class frequency. Each picture will represent the unit of 10, 20, 100, 1000, 10,000, lacks etc.

e.g. Production of car per month.

MAP DIAGRAM (SPOT DIAGRAM)

Represents the geographical distribution of frequencies of frequencies of a variable / characteristics.

e.g. IMR of South India.



CHARACTERISTICS OF FREQUENCY DISTRIBUTION

Measures of location

Major characteristics of frequency distribution are –

- Measures of Central tendency (Location, Position, Average)
- Measures of scatteredness / Degree of scatteredness (Dispersion, / Variability / Spread)
- Extent of symmetry – If the data are asymmetrical called as **"Skewness,"** which can be of two types –
  - ✓ Positive Skewness (Right sided)
  - ✓ Negative Skewness (Left sided)
- Measures of Peakedness – If it is abnormally peak or flat is called as **"Kurtosis."**

MEASURES OF CENTRAL TENDENCY

It is one among the characteristic of frequency distribution.

Definition

It refers to a single central number or value that condenses the mass data and enables us to give an idea about the whole or entire data.

The commonly used measures of central tendencies are –

01. Arithematic mean ($\overline{x}$)
02. Median (Q_2)
03. Mode (z)

A good measure of central tendency should posses the following properties –

- ✯ Easy to understand.
- ✯ Easy to calculate.
- ✯ Based on all observations.
- ✯ Should be properly defined.
- ✯ Should be used for further mathematical calculations.
- ✯ Should not be affected by extreme high or low values.

SELECTION OF CENTRAL TENDENCY

If the distribution is symmetrical one should select the Arithmetic Mean and if the distribution is Skewness (Asymmetry) one should use either median or mode.

MEAN / ARITHMETIC MEAN / AVERAGE

Introduction

It is a most preferred and commonly used measure of central tendency.

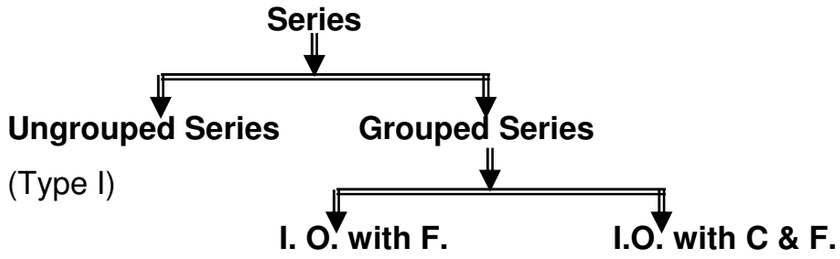
It is also called as **"Average."**

Definition

It means, the additional / summation of all individual observations divided by total number of observations.

Types of Series / Problems

There are 2 types of series –



[Where, I. O. – Individual Observation, F – Frequency, C – Class.]

- Ungrouped Series – Includes individual observations without frequency.
- Grouped Series – Includes individual observations with frequency and class frequency.

CALCULATION FOR TYPE I SERIES –

(Individual Observation without frequency)

Direct Method (DM)

Formula = $\bar{X} = \Sigma x / n$

Where, $\bar{X}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, n – is Total number of observations.

Step Deviation Method (SDM) or Indirect method

Formula = $\bar{X} = A + \Sigma d / n$ (Where, $d = x - A$.)

Where, $\bar{X}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, A – is assumed value, d – deviated value, n – is Total number of observations.

e.g. Following is the data showing the Montaux test of 6 children.

2, 4, 7, 3, 5, 6.

The arithmetic mean of the above given set of data can be calculated by 2 methods –

- Direct Method
- Step Deviation Method

DIRECT METHOD

Formula = $\bar{x} = \Sigma x / n$

Where, $\bar{x}$ – is Arithmetic mean, Σ – is Summation of all observations, x – is individual observation, n – is Total number of observations.

$$\bar{x} = \frac{2 + 4 + 7 + 3 + 5 + 6.}{6}$$

$\bar{x} = 27 / 6 = 4.5$

So, the Arithmetic mean of the above given data is 4.5.

STEP DEVIATION METHOD

Formula :: $\bar{x} = A + \Sigma d / n$ (Where, $d = x – A.$)

Where, $\bar{x}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, A – is assumed value, d – deviated value, n – is Total number of observations.

Step 1st : Calculate d. (i.e. Deviated value)

It is calculated by $d = x – A$.

Consider A – is 10. (i.e. Assumed value.)

$x – A = d$

$2 – 10 = – 8$

$4 – 10 = – 6$

$7 – 10 = – 3$

$3 – 10 = – 7$

$5 – 10 = – 5$

$6 – 10 = – 4$

Step 2nd : Calculate summation of d

Summation = (– 8) + (– 6) + (– 3) + (– 7) + (– 5) + (–4)

= – 33.

Step 3rd : Calculate Arithmetic mean.

$\bar{x} = 10 + (– 33) / 6$

$\bar{x} = 10 + (– 5.5) = 4.5.$

So, the arithmetic mean of the above given data is 4.5 calculated by SDM.

CALCULATION FOR TYPE II SERIES –

(Individual Observation with frequency)

Direct Method (DM)

Formula = $\bar{x} = \Sigma f x / n$

Where, $\bar{x}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, n – is Total number of observations, f – Individual frequency, x – Individual observation.

Step Deviation Method (SDM)

Formula = $\bar{x} = A + \Sigma f d / n$ (Where, $d = x - A$.)

Where, $\bar{x}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, A – is assumed value, d – deviated value, n – is Total number of observations, f– Individual frequency, x – Individual Observation

e.g. The number of children in family for 50 couples are as follows –

Number of children (x)	Number of couples (f)	f x
0	4	0
1	9	9
2	10	20
3	12	36
4	7	28
5	6	30
6	2	12

The arithmetic mean of the above given set of data can be calculated by 2 methods –

- Direct Method
- Step Deviation Method

DIRECT METHOD

Formula = $\bar{x} = \Sigma fx / n$

Where, $\bar{x}$ – is Arithmetic mean, Σ – is Summation of all observations, x–is individual observation, n– Total number of observations, f- Frequency

$\bar{x} = \frac{135.}{50}$

$\bar{x}$ = 2.7 i.e. Approximately 3 children per family.

So, the Arithmetic mean of the above given data is 2.7 i.e. 3.

STEP DEVIATION METHOD

Formula = $\bar{X} = A + \Sigma fd / n$ (Where, $d = x - A$.)

Where, $\bar{X}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, A – is assumed value, d – deviated value, n – is Total number of observations, x – Individual observation.

Step 1st : Calculate d and fd.

It is calculated by d = x – A. (i.e. Deviated value)

Consider **A is 3**. (i.e. Assumed value.)

x – A = d	**= fd**
0 – 3 = – 3 x 4	= – 12.
1 – 3 = – 2 x 9	= – 18
2 – 3 = – 1 x 10	= – 10
3 – 3 = 0 x 12	= 0
4 – 3 = 1 x 7	= 7
5 – 3 = 2 x 6	= 12
6 – 3 = 3 x 2	= 6

Step 2nd : Calculate summation of fd

Summation = (– 12) + (– 18) + (– 10) + (0) + (7) + (12) + (6)

= – 15.

Step 3rd : Calculate Arithmetic mean.

$\bar{X}$ = 3 + (– 15) / 50

$\bar{X}$ = 3 + (– 0.3) = **2.7.**

So, the arithmetic mean of the above given data is 2.7 calculated by SDM.

CALCULATION FOR TYPE III SERIES –

(Individual Observation with class and frequency)

Direct Method (DM)

Formula = $\bar{X} = \Sigma f x / n$

Where, $\bar{X}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, n – is Total number of observations, f – Class Frequency, x – Class midpoint.

Step Deviation Method (SDM)

Formula = $\bar{X} = A + \Sigma f d / n$ (Where, $d = x - A$.)

Where, $\bar{X}$ – is Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, A – is assumed value, d – deviated value, n – is Total number of observations, f – Class frequency, x – Class midpoint.

e.g. Following are the waiting time of 20 patients to consult a physician in clinic –

Class	Frequency	Class midpoint (x)	fx
0 – 5	3	2.5	7.5
5 – 10	2	7.5	15
10 – 15	3	12.5	37.5
15 – 20	5	17.5	87.5
20 – 25	3	22.5	67.5
25 – 30	4	27.5	110
			325

The arithmetic mean of the above given set of data can be calculated by 2 methods –

- Direct Method
- Step Deviation Method

DIRECT METHOD

Formula = $\bar{x} = \Sigma fx / n$

Where, $\bar{x}$ – is Arithmetic mean, Σ – is Summation of all observations, x – is class midpoint, n – is Total number of observations, f – Class frequency.

$$\bar{x} = \frac{325}{20}$$

$\bar{x}$ = 16.25 i.e. Approximately 17 minutes per minutes.

So, the Arithmetic mean of the above given data is 16.25 i.e. 17.

STEP DEVIATION METHOD

Formula = $\bar{x} = A + \Sigma fd / n$ (Where, d = x – A.)

Where, $\bar{x}$ – Arithmetic mean, Σ – is Sigma (i.e. Summation of all observations, A – Assumed value, d – deviated value, n – is Total number of observations, x – Class mid point, f – Class frequency.

Step 1st : Calculate d and fd.

It is calculated by d = x – A. (i.e. Deviated value)

Consider A – is 15. (i.e. Assumed value.)

x – A = d	= fd	
2.5 – 15	= – 12.5 x 3	= – 37.5.
7.5 – 15	= – 7.5 x 2	= – 15.
12.5 – 15	= – 2.5 x 3	= – 7.5.
17.5 – 15	= 2.5 x 5	= 12.5.
22.5 – 15	= 7.5 x 3	= 22.5.
27.5 – 15	= 12.5 x 4	= 50.

Step 2nd : Calculate summation of fd.

Summation = (– 37.5) + (– 15) + (– 7.5) + (12.5) + (22.5) + (50)

= (– 60) + 85.

= 25.

Step 3rd : Calculate Arithmetic mean.

$\bar{x}$ = 15 + (25) / 20

$\bar{x}$ = 15 + (1.25) = 16.25.

So, the arithmetic mean of the above given data is 16.25 calculated by SDM. i.e. Approximately 17 minutes a patient should wait to consult to physician in clinic.

MATHEMATICAL PROPERTIES OF ARITHMETIC MEAN

01. The sum of the deviation from the arithmetic mean is always zero for a given distribution.

$$\text{i.e. } \mathcal{E}(x - \bar{x}) = 0.$$

Where, x – Individual observation, $\bar{x}$ – Arithmetic mean, $\mathcal{E}$– Summation.

It is because of this property the mean is characterized as a point of balance. i.e. **"The sum of the positive deviation of the mean is exactly equal to the negative deviation of the mean."**

e.g. Weight of 6 students are – 10 kg, 12 kg, 11 kg, 14 kg, 15 kg, 13 kg each.

Arithmetic mean of the above mentioned set of data is as follows –

Formula = $\bar{x} = \mathcal{E}x / n$

Where, $\bar{x}$ – is Arithmetic mean, $\mathcal{E}$ – is Summation of all observations,

x – is individual observation, n – is Total number of observations.

$$\bar{x} = \frac{10 + 12 + 11 + 14 + 15 + 13}{6}.$$

$\bar{x}$ = 75 / 6 = 12.5.

So, the Arithmetic mean of the above given data is 12.5.

i.e. $(x - \bar{x}) = 0.$

10 – 12.5 = – 2.5.

12 – 12.5 = – 0.5

11 – 12.5 = – 1.5

14 – 12.5 = 1.5

15 – 12.5 = 2.5

13 – 12.5 = 0.5 = **0.**

Summation of the $\mathcal{E}(x - \bar{x}) = 0.$

02. COMBINED ARITHMETIC MEAN

It can be calculated out of Arithmetic means of several sets of data.

e.g. For 2 sets of data combined arithmetic mean will be as follows –

$$\textbf{CAM} :: \bar{X}_{1,2} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

e.g. A student has scored 60% marks in SSLC and 70% in PUC with 6 subjects each. Calculate the combined Arithmetic Mean.

Here, $n_1=6$, $\bar{X}_1=60$, $n_2=6$, $\bar{X}_2=70$

$$\textbf{CAM} :: \bar{X}_{1,2} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

$$\bar{X}_{1,2} = \frac{6 \times 60 + 6 \times 70}{6+6}$$

$$= \frac{360+420}{12}$$

$$= \frac{780}{12}$$

$$\bar{X}_{1,2} = 65\%$$

03. WEIGHTED OF ARITHMETIC MEAN

It is based on weighted or importance.

Arithmetic Mean gives equal importance to all observations, but in some cases, all the observations do not have same importance. When this is true, weighted Arithmetic Mean is calculated.

It enables to calculate an average that takes into account, the importance of each value to the overall total.

It is calculated by,

$$\bar{X}_w = \frac{\Sigma wx}{\Sigma w}$$

Where, W= weighted given to each observation, $\bar{X}_w$ = Weighted Arithmetic Mean, Σ - is summation, x – is individual observation.

e.g. If a student scores following marks in 3 examination taking into consideration. Viz. –

Exams	Weighted	Marks scored	wx
1st Exam	25%	60	1500
2nd Exam	25%	30	750
3rd Exam	50%	90	4500
			6750

Respective percentages are – 60, 30, 90. Calculate weighted of Arithmetic mean.

It is calculated by,

$$\bar{x}_w = \frac{\varepsilon wx}{\varepsilon w}$$

Where, W= weighted given to each observation, $\bar{x}_w$ = Weighted Arithmetic Mean, ε - is summation, x – is individual observation.

= 6750 / 100 = 67.5%.

So, the weighted arithmetic mean is 67.5%.

MERITS

- It is correctly / rigidly defined.
- Easy to understand.
- Easy to calculate.
- Based on each and every observation.
- Very familial concept to the people.
- Every set of data will have Arithmetic mean.
- Every set of data has one and only one Arithmetic mean.
- Used for further mathematical calculations like – Standard deviation.

DEMERITS

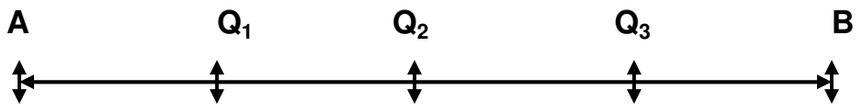
- Affected by extreme values (either low / high).
- Cannot be detected by mere inspection of the data.
- It can not be obtained even if a single value is missing.
- It can not be used for qualitative data.

MEDIAN (Q_2)

It is called Q_2 because it denotes 2nd Quartile or positional value.

Introduction

It is the 2nd measure of central tendency. Here there are 3 quartiles Q_1, Q_2, Q_3 which divides the distribution into 4 parts or equals.



Definition

Median or 2nd quartile (Q_2) divides the distribution into two equal parts i.e. 50% of the distribution is below the median & 50% is above the median.

Q_1 = n / 4. & Q_3 = 3 x n / 4 item. Where, n – is total number of observations.

CALCULATION

Type I Problem

A) When 'n' is odd (n – Total number observation)

If the total number of observations are odd, then arrange the observations either in ascending or descending order and calculate the median by following method –

$$Q_2 = \frac{n+1}{2} \text{ item}$$

Where, Q_2 – is median and n – is total number of observations

e.g. Number of patients treated in emergency room on 7 consecutive days are as 86, 49, 52, 43, 25, 11, 31. Calculate the median.

Answer :

Arranging the observations in ascending order –

11, 25, 31, 43, 49, 52, 86

Total number of observations are 7. i.e. Odd number.

So,
$$Q_2 = \frac{n+1}{2} \text{ item}$$

Where, Q_2 – is median and n – is total number of observations

$Q_2 = 7 + 1 / 2$

$Q_2 = 8 / 2$

$Q_2 = 4^{th}$ item. i.e. **43.**

So, the median of above given set of data is 43. (i.e. 4th item)

B) <u>When 'n' is even</u> (n – Total number observation)

If the total numbers of observations are even, then the median is the average of two meddle items after they have been arranged in ascending or descending order.

$$Q_2 = \frac{A+B}{2}$$

Where, Q_2 – is median and A & B – are the 2 middle items in a given set of data.

e.g. The number of patients treated in OPD treated for 6 consecutive days –
11, 12, 10, 31, 34, 30. Then calculate median.

Answer :

Arranging the observations in ascending order –

10, 11, 12, 30, 31, 34. Where A is equal to 12 & B is equal to 30.

Total number of observations are 6. i.e. Even number.

$$Q_2 = \frac{A+B}{2}$$

Where, Q_2 – is median and A & B – are the 2 middle items in a given set of data.

$Q_2 = 12 + 30 / 2$

$Q_2 = 42 / 2$

$Q_2 = 21.$

So, the median of above given set of data is 21.

<u>Type II Problem</u>

A cumulative frequency distribution table is constructed.

n / 2 item is calculated and identified in CFD (Cumulative Frequency Distribution) and the median the corresponding x value of n / 2 item.

e.g. Table showing number of illness in a patients.

No. of Illness (x)	Frequency (f) No. of patients	CFD	
0	24	24	
1	76	100	
2	114	214	
3	**115**	**329**	**Q2**
4	86	415	
5	57	472	
6	26	498	
7	18	516	

Q_2 = n / 2 item. (Calculation of Median for even number of observations)

Where, Q_2 – is Median, n – is total number of observations.

$Q_2 = 516 / 2 = 258^{th}$ item.

Identify the 258^{th} item in CFD (i.e. 329) is the median and the corresponding x value is the median. (i.e. 3)

i.e. **The median is 3.**

Type III problem

Median class should be identified by using cumulative frequency distribution. **i.e. Q_2 = n / 2 value.** The various related values are identified and calculated.

Formula = $Q_2 = L_1 + \frac{L_2 - L_1}{f}(q_2 - pcf)$

Where, Q_2 – is Median, L_1 – is Lower limit of Median class.

L_2 – is Upper limit of Median class, pcf – is Preceding Cumulative Frequency (i.e. Previous / preceding CF of Median class.)

f – Frequency of the Median class.

e.g. Following table showing expenditure of the 1000 individuals in the age group of 20 to 60 years.

Age	Frequency	Cumulative frequency distribution	
20 – 25	120	120	
25 – 30	125	245	
30 – 35	180	425	
35 – 40	**160**	**585**	**Q_2**
40 – 45	150	735	
45 – 50	140	875	
50 – 55	100	975	
55 – 60	25	1000	

Median i.e. Q_2 = n / 2 (Calculation of Median for even number of observations)

Q_2 = 1000 / 2 = 500.

Formula = $Q_2 = L_1 + \frac{L_2 - L_1}{f} \times (q_2 - pcf)$

Where, Q_2 – is Median, L_1 – is Lower limit of median class, L_2 – is Upper limit of Median class, f – is frequency of median class, q_2 – is ½ of the total number of observations, pcf – Preceding cumulative frequency, CF – is cumulative frequency.

Q_2 $= 35 + \frac{40 - 35}{160}(500 - 425)$

$= 35 + \frac{5 \times 75}{160}$

$= 35 + 375 / 160 = 35 + 2.4 = 37.34$

Q_2 = 37.34.

Merits

- Easy to understand.
- Easy to calculate.
- Not affected by extreme values.
- Only average to be used dealing with the qualitative data.
- Used to determine the typical values.
- Merely by inspection, median can be calculated in some cases only.

De-merits

- Mode is not based on the all the observations. (i.e. Gives only positional values)
- Not used for further mathematical calculations.
- In case of even numbers of observations median can be determined exactly.

MODE (z)

Dictionary meaning of the mode is common, fashionable or usual. Mode is the value which occurs more frequently (i.e. Maximum number of times) in a given set of data and around which other items of the set cluster each other (i.e. Central point of alteration)

Type I :

Selection of Mode = The Observation having highest repetition.

Find out the mode of the following data.

10, 11, 12, 26, 20, 40, 20, 10, 12, 10.

As 10 is repeating 3 times 10 is the mode.

But, some times there can be no mode (i.e. 1, 2, 3, 4, 5, 6.) or more than one mode (i.e. 1, 1, 1, 2, 2, 3, 3, 4, 4, 4, 4.).

Type II :

Selection of Mode Observation = Observation containing highest frequency.

Following table showing number of children per family.

Number of children per family	Number of families
0	13
1	24
2	25
3	13
4	14

In this case, the data which has maximum frequency is taken as Mode (z).

In the above series the observations which has maximum frequency is the mode. As 2 has maximum frequency i.e. 25.

Hence, the mode of the above given set of data is 2.

Type III :

Selection of Model class = The class containing highest frequency

Formula = $\text{Mode (z)} = \dfrac{L_1 + f_1 - f_0}{2f_1 - f_0 - f2} \times C$

Where, z – is Mode, L_1 – is the lower limit of the modal class,

f_1 – is frequency of the modal class, f_0 – is frequency of previous class,

f_2 – is frequency of next class, c – is class interval.

If the modal class is 1st or last class their frequencies f_0 & f_2 should be taken as 0.

e.g. Following Table showing the Age wise distribution of 150 patients.

Age groups		Frequency (f)	
20 – 30		15	
30 – 40		23	
40 – 50		27	
l_0	50 – 60	20	f_0
l_1	60 – 70	35	f_1
l_2	70 – 80	25	f_2
80 – 90		5	

Formula = Mode (z) = $L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$

Where, z – is Mode, L_1 – is the lower limit of the class model,

f_1 – is frequency of the modal class, f_0 – is frequency of previous class,

f_2 – is frequency of next class, c – is class interval.

Mode (z) = $60 + \frac{35 - 20}{2 \times 35 - 20 - 25} \times 10.$

= $60 + \frac{15}{70 - 20 - 25.} \times 10$

= $60 + \frac{15 \times 10}{50 - 25.}$

= $60 + \frac{150}{25}$

= 60 + 6

= 66.

Mode (z) = 66.

Merits

- Most representative value of a given set of data.
- Easy to calculate.
- Not affected by extreme values.
- Mode can be found for both qualitative and quantitative data.
- Easy to understand.
- Average to be used to find the ideal size.

De-merits

- Sometimes no mode or more than one mode in a given set of distribution.
- Not used for further mathematical calculations.
- Not commonly used.

DISPERSION or VARIATION

MEASURES OF VARIABILITY

Introduction

In the previous chapter on measure of central tendency, it was providing us a single representation value of a given set of data. But that alone may not be adequate to describe the complete data.

e.g. Table showing marks scored by the 3 students in 6 subjects.

Subjects/ Students	A	B	C
1st Subject	50	49	80
2nd Subject	50	51	20
3rd Subject	50	48	60
4th Subject	50	52	40
5th Subject	50	47	70
6th Subject	50	53	30

The arithmetic mean of all the above students are same i.e. 50. But student A has no variation. Student B has little variation and student C has more variation. This scatteredness can be calculated by various measures of variability / dispersion.

Definition

Measures of variation / dispersion describe the spread or scatteredness of the individual observations or items around the central tendency.

Significance

- Gives complete idea / picture of data.
- Gives information about scatteredness around the central tendency.
- Useful for further calculations e.g. Test of significance, etc.
- Helps in comparison of distribution.
- Gives idea about the reliability of average value.

Methods of Dispersion

Commonly used methods are –

- Range
- Inter quartile range (IQR)
- Semi inter-quartile Range / Quartile deviation (QD)
- Mean deviaiton / Average deviation (MD)
- Standard deviation (SD)

RANGE

Range is defined as the difference between the highest and lowest values in a set of data.

Calculation

R = H – L

Where, R = Range, H = Highest value, L = Lowest value.

e.g. Following is the Hb% of 6 children. Calculate the range.

8.8 gm%, 9.3 gm%, 10.5 gm%, 11.4 gm%, 14 gm%, 10.5 gm%.

Formula – R = H – L

Where, R = Range, H = Height value, L = Lowest value.

R = 14 – 8.8 = 5.2 R = 5.2.

So, the range of Hb% of 6 children is 5.2.

Relative Measure Of Range

It is also called as coefficient of range.

Co-efficient of $R = \frac{H - L}{H + L} \times 100$

Where, R – is Range, H – is Highest value, L – is Lowest value.

$$R = \frac{14 - 8.8}{14 + 8.8} \times 100$$

$$R = \frac{5.2}{22.2} \times 100$$

$$R = 23.42\ \%$$

Coefficient of Range (R) = 23.42%.

Merits

- Easy to understand and calculate.
- Easy to compare.
- Gives first hand information about variation.

De-merits

- It is not based on all the values.
- Affected by extreme values.

INTER QUARTILE RANGE (IQR)

It is defined as the difference between the 3rd quartile and 1st quartile.

Formula = IQR = $Q_3 - Q_1$.

Where, Q_3 – is 3rd quartile = 3 x n / 4. Q_1 – is 1st quartile = n / 4.

n – is Number of observations.

e.g. Following are the weights of 10 students. Calculate the IQR.

84 Kg., 48 Kg., 39 Kg., 64 Kg., 78 Kg., 63 Kg., 38 Kg., 54 Kg., 60 Kg., 62 Kg.

Ascending order –

38, 39, 48, 54, 60, 62, 63, 64, 78, 84. (Even numbers method for Median)

Formula = Q_3 = 3 x n / 4.

Where, Q_3 – is 3rd quartile = 3 x n / 4. n – is Number of observations.

Q_3 = 3 x 10 / 4 = 7.5 = 8th item is 64.

Formula = Q_1 = n / 4.

Where, Q_1 – is 1st quartile = n / 4. n – is Number of observations.

Q_1 = 10 / 4 = 2.5 = 3rd item is 48.

Formula = IQR = $Q_3 - Q_1$.

Where, IQR – is Inter quartile range, Q_3 – is 3rd quartile, Q_1 – is 1st quartile

IQR = 64 – 48

IQR = 16.

Merits

- Simple and easy to understand.
- Easy to calculate.
- Not affected by extreme values.

De-merits

- It is a positional value, which is based on 2 quartiles.
- Based on first and last values. (i.e. Initial and last 25% values are not included)

SEMI-QUARTILE RANGE OR QUARTILE DEVIATION

It is a measure of variability.

It is calculated by the average difference of 3rd quartile and 1st quartile.

Formula = QD = IQR / 2. = $Q_3 – Q_1$ / 2.

Where, QD – is Quartile Deviation, IQR – is Inter quartile deviation.

Q_3 – is Item 3rd quartile, Q_1 – is Item 1st quartile

e.g. Following are the weights of 10 students. Calculate the IQR.

84 Kg., 48 Kg., 39 Kg., 64 Kg., 78 Kg., 63 Kg., 38 Kg., 54 Kg., 60 Kg., 62 Kg.

Ascending order –

38, 39, 48, 54, 60, 62, 63, 64, 78, 84. (Even numbers method for Median)

Q_3 – is 64.

Q_2 – is 48.

IQR – is 16.

Formula = QD = IQR / 2. = $Q_3 – Q_1$ / 2.

Where, QD – is Quartile Deviation, IQR – is Inter quartile deviation.

Q_3 – is Item 3rd quartile, Q_1 – is Item 1st quartile

QD = 16 / 2

QD = 8.

Coefficient of QD = $\frac{Q_3 – Q_1}{Q_3 + Q_1} \times 100$

Where, RD Range deviation, Q_3 – is 3rd quartile, Q_1 – is 1st quartile.

e.g. Following are the weights of 10 students. Calculate the IQR.

84 Kg., 48 Kg., 39 Kg., 64 Kg., 78 Kg., 63 Kg., 38 Kg., 54 Kg., 60 Kg., 62 Kg.

Ascending order –

38, 39, 48, 54, 60, 62, 63, 64, 78, 84. (Even numbers method for Median)

Q_3 – is 64.

Q_2 – is 48.

Coefficient of RD = $\frac{64 – 48}{64 + 48} \times 100$

= 16 / 112 x 100

Coefficient of RD = 14.28 %

Merits

- Easy and simple to understand.
- Easy to calculate.
- Not affected by extreme values.

Demerits

- It is a positional valve which is based on two quartiles.
- Based on 1st & last value (First 25% and last 25% are not included.)

MEAN DEVIATION / AVERAGE DEVIATION (MD/AD)

Introduction

It is the improvement of previous methods of variation. Because, it considers all the observations in a given set of data.

Definition

It is an average amount of scatter of the items in a distribution from any measure of the central tendency. (i.e. May be Mean, Mode, etc.) by ignoring the mathematical signs.

Calculations

It is calculated by –

Formula = $AD = \frac{\mathcal{E}|x - \bar{x}|}{n}$

Where, AD – is Average Deviation / Mean deviation, $\mathcal{E}$ – is summation, | | – is Modulus, x – is Individual observation, $\bar{x}$ – is Arithmetic mean, n – is Total number of observations.

e.g. Number of students in a single class in different divisions.

10, 20, 30, 40, 50. Calculate the Average Mean.

Step 1st : Calculate the arithmetic mean.

Formula = $\bar{x} = \frac{\mathcal{E}\, x}{n}$

Where, $\bar{x}$ – is Arithmetic mean, $\mathcal{E}$ – is summation, x – is individual observations, n – is total number of observations.

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50}{5}$$

$$\bar{x} = \frac{150}{5}$$

$$\bar{x} = 30.$$

Step 2nd : Calculate the Average Mean.

It is calculated by –

Formula = $AD = \frac{\mathcal{E}|x - \bar{x}|}{n}$

Where, AD – is Average Deviation / Mean deviation, $\mathcal{E}$ – is summation, | | – is Modulus, x – is Individual observation, $\bar{x}$ – is Arithmetic mean, n – is Total number of observations.

Calculate x – $\bar{x}$

10 – 30 = – 20, 20 – 30 = – 10, 30 – 30 = 0, 40 – 30 = 10, 50 – 30 = 20.

$$AD = \frac{\mathcal{E}|x - \bar{x}|}{n}$$

$$AD = \frac{\mathcal{E}|20 + 10 + 0 + 10 + 20|}{5}$$

AD = 60 / 5 = **12.**

So, the absolute average deviation of the given set of the data is 16.

Relative / Co-efficient of Average Deviation

Formula = CAD = AD / Mean x 100.

Where, CAD – is Coefficient of Average Deviation,

AD – Average Deviation, Mean – is Arithmetic Mean.

CAD = 12 / 30 x 100

CAD = 0.4 x 100

CAD = 40 %.

Merits

- Easy to calculate.
- Easy to Understand.
- Based on all the observations.

De-merits

- Ignore the mathematical signs. Because, if it does not ignore the mathematical signs then, sum of deviation from the arithmetic mean will be zero. **(i.e. $\mathcal{E}(x - \bar{x} = 0)$**

STANDARD DEVIATION (SD)

Introduction

It is a most widely used and the best method of calculating deviation.

While calculating the Average deviation (AD), though it takes into consideration of all the observations, it ignores the mathematical signs. But, standard deviation (SD) overcomes this problem by squaring the deviation.

Definition

The Standard deviation is the square root of summation of square of deviation of given set of observations from the arithmetic mean divided by the total number of observations.

Calculations

It is calculated by following ways –

Type I : Individual observation without frequency.

Formula $= \sigma = \sqrt{\frac{\varepsilon (x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of,
x – is Individual observation, $\bar{x}$ – is Arithmetic mean,
n – is Total number of observation.

e.g. Following are the results of the ESR in mm for 1st hour observed in 5 individuals. Calculate the standard deviation.

2, 4, 6, 8, 10.

The above mentioned example comes under the Type I series of data

Step 1st : Calculate the arithmetic mean.

Formula = $\bar{x} = \frac{\varepsilon x}{n}$

Where, $\bar{x}$ – is Arithmetic mean, ε – is summation,
x – is individual observations, n – is total number of observations.

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5}$$

$$\bar{x} = \frac{30}{5}$$

$$\bar{x} = 6.$$

Step 2nd : Calculate the $x - \bar{x}$.

$2 - 6 = -4$

$4 - 6 = -2$

$6 - 6 = 0$

$8 - 6 = 2$

$10 - 6 = 4$

Step 3rd : Calculate the summation of $(x - \bar{x})^2$.

$(-4)^2 = 16$

$(-2)^2 = 4$

$(0)^2 = 0$

$(2)^2 = 4$

$(4)^2 = 16$

$16 + 4 + 0 + 4 + 16 = 40$.

Step 4th : Calculate the Standard Deviation (SD).

The above mentioned example comes under the Type I series of data.

Formula $= \sigma = \sqrt{\frac{\varepsilon (x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of,

x – is Individual observation, $\bar{x}$ – is Arithmetic mean,

n – is Total number of observation.

$$\sigma = \sqrt{\frac{40}{5}}$$

$\sigma = 2.8$.

So the standard deviation of the above given set of data is 2.8.

Coefficient of Standard Deviation / Coefficient of Variation (CSD/CV)

Formula = Coefficient of Variation = SD / AM x 100

Where, SD – is Standard Deviation, AM – is Arithmetic Mean.

CSD / CV = 2.8 / 5 x 100

CV = 47%.

Type II : Individual observation with frequency.

Formula $= \sigma = \sqrt{\frac{\varepsilon f(x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of,

F – is frequency, x – is Individual observation, $\bar{x}$ is Arithmetic mean,

n – is Total number of observation.

e.g. Following table shows the number of children per family. Calculate the standard deviation.

Number of Children (x)	Number of families (f)
1	2
2	3
3	2
4	4
5	3
Total Number of Observations = 14. (Add all frequencies)	

Step 1st : Calculate the arithmetic mean.

Formula = $\bar{x} = \frac{\Sigma fx}{n}$

Where, $\bar{x}$ – is Arithmetic mean, Σ – is summation, f – is frequency, x – is individual observations, n – is total number of observations.

$\bar{x} = \frac{((2x1) + (3x2) + (2x3) + (4x4) + (3x5))}{5}$

$\bar{x} = \frac{45}{14}$

$\bar{x} = 3.21.$

Step 2nd : Calculate the $x - \bar{x}$.

1 – 3.21 = – 2.21

2 – 3.21 = – 1.21

3 – 3.21 = 0.21

4 – 3.21 = 0.79

5 – 3.21 = 1.79

Step 3rd : Calculate the summation of $f (x - \bar{x})^2$.

$(– 2.21)^2$ = 4.88 x 2 = 9.76.

$(– 1.21)^2$ = 1.46 x 3 = 4.38.

$(0.21)^2$ = 0.04 x 3 = 0.08.

$(0.79)^2$ = 0.62 x 4 = 2.48.

$(1.79)^2$ = 3.20 x 3 = 9.6.

Summation = 9.76 + 4.38 + 0.08 + 2.48 + 9.6

$\Sigma f(x - \bar{x})^2 = 26.3.$

Step 4th : Calculate the Standard Deviation (SD).

Formula $= \sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of, F – is frequency, x – is Individual observation, $\bar{x}$ – is Arithmetic mean, n – is Total number of observation.

$\sigma = \sqrt{26.3 / 14}$

$SD = \sqrt{1.87}$

SD = 1.36.

So, the standard deviation of the given set of data is 1.36.

Coefficient of Standard Deviation / Coefficient of Variation

Formula = CV = SD / AM x 100.

Where, CV – Coefficient of Variation, SD – is Standard Deviation, AM – is Arithmetic mean.

CV = 1.36 / 3.21 x 100.

Coefficient of Variation = 42.36%.

Type III : Class and frequency.

Formula $= \sigma = \sqrt{\frac{\varepsilon f(x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of,

F – is frequency, x – is Class midpoint, $\bar{x}$ – is Arithmetic mean,

n – is Total number of observation.

e.g. Following are the number of patients according to the age groups. Calculate the standard deviation.

Sl.	Age groups	No. of Pt.'s
01.	10 – 20	2
02.	20 – 30	1
03.	30 – 40	3
04.	40 – 50	4
Total number of Observations		**10**

Step 1st : Calculate the arithmetic mean.

Formula = $\bar{x} = \frac{\varepsilon fx}{n}$

Where, $\bar{x}$ – is Arithmetic mean, ε – is summation, f – is frequency, x – is class mid point, n – is total number of observations.

Class midpoint

Formula = C.M. = L.L. + U.L. / 2

Where, C. M. – is Class midpoint, L.L. – is Lower limit of the class,

U.L. – is Upper limit of the class.

CM = L.L. + L.L. / 2

10 + 20 / 2 = 15.

20 + 30 / 2 = 25.

30 + 40 / 2 = 35.

40 + 50 / 2 = 45.

$\bar{x} = \frac{((2x15) + (1x25) + (3x35) + (4x45))}{10}$

$\bar{x} = \frac{340}{10}$

$\bar{x} = 34.$

Step 2nd : Calculate the $x - \bar{x}$.

15 – 34 = – 19

25 – 34 = – 9

35 – 34 = 1

45 – 34 = 11

Step 3rd : Calculate the summation of $f (x - \bar{x})^2$.

$(-19)^2$ = 361 x 2 = 722.

$(-9)^2$ = 81 x 1 = 81.

$(1)^2$ = 1 x 3 = 3.

$(11)^2$ = 121 x 4 = 484.

Summation = 722 + 81 + 3 + 484

$\varepsilon f(x - \bar{x})^2 = 1290.$

Step 4th : Calculate the Standard Deviation (SD).

Formula $= \sigma = \sqrt{\frac{\varepsilon f(x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of,

F – is frequency, x – is class midpoint $\bar{x}$ - is Arithmetic mean,

n – is Total number of observation.

$\sigma = \sqrt{1290 / 10}$

$SD = \sqrt{129}$

SD = 11.35.

So, the standard deviation of the given set of data is 1.36.

Coefficient of Standard Deviation / Coefficient of Variation

Formula = CV = SD / AM x 100.

Where, CV – Coefficient of Variation, SD – is Standard Deviation,

AM – is Arithmetic mean.

CV = 11.35 / 34 x 100.

Coefficient of Variation = 38.59 %.

Significance

- Based on all observations.
- Best method of calculation without ignoring mathematical signs.
- Useful for further statistical calculations. (i.e. Test of Significance, etc.)
- Useful for calculation of standard error.
- Lesser the standard deviation, better the estimation of population mean.

STANDARD ERROR (SE)

Introduction

In medical investigations only a sample portion of the population is studied. The sample results are bounded to differ from population results.

This difference or the error is measured by **"Standard Error."**

The word error here means – **"The difference between the true value of a population parameter and estimated value provided by appropriate sample statistics."**

Definition

The standard error of the mean is the – **"Standard deviation of the sample mean divided by the square root of the sample size."**

Formula $SE = SD / \sqrt{n}$

Where, SE – is Standard error, SD – is Standard Deviation,

n – is the total number of observations.

Calculation

e.g. Following are the results of the ESR in mm for 1st hour observed in 5 individuals. Calculate the standard error.

2, 4, 6, 8, 10.

The above mentioned example comes under the Type I series of data

Step 1st : Calculate the arithmetic mean.

Formula = $\bar{x} = \frac{\Sigma x}{n}$

Where, $\bar{x}$ – is Arithmetic mean, Σ – is summation,
x – is individual observations, n – is total number of observations.

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5}$$

$$\bar{x} = \frac{30}{5}$$

$$\bar{x} = 6.$$

Step 2nd : Calculate the $x - \bar{x}$.

$2 - 6 = -4$

$4 - 6 = -2$

$6 - 6 = 0$

$8 - 6 = 2$

$10 - 6 = 4$

Step 3rd : Calculate the summation of $(x - \bar{x})^2$.

$(-4)^2 = 16$

$(-2)^2 = 4$

$(0)^2 = 0$

$(2)^2 = 4$

$(4)^2 = 16$

$16 + 4 + 0 + 4 + 16 = 40$.

Step 4th : Calculate the Standard Deviation (SD).

The above mentioned example comes under the Type I series of data.

Formula $= \sigma = \sqrt{\frac{\varepsilon (x - \bar{x})^2}{n}}$

Where, σ – is Standard Deviation, ε – is Summation of,
x – is Individual observation, $\bar{x}$ – is Arithmetic mean,
n – is Total number of observation.

$$\sigma = \sqrt{\frac{40}{5}}$$

$\sigma = 2.8$.

So the standard deviation of the above given set of data is 2.8.

Step 5th : Calculate Standard Error

The standard error of the mean is the Standard deviation of the sample mean divided by the square root of the sample size.

Formula $SE = SD / \sqrt{n}$
Where, SE – is Standard error, SD – is Standard Deviation,
n – is the total number of observations.

$SE = 2.8 / \sqrt{5}$

$SE = 2.8 / 2.23$

$SE = 1.25$.

So, the Standard error of the given set of data is 1.25.

Interpretation

- The value of the standard error (SE) is directly proportional with the standard deviation (SD). i.e. Higher the SD, higher the SE.

SE α SD
Where, SE – is the Standard error, SD – is Standard deviation.

- The value of the standard error (SE) is inversely proportional with the sample size. i.e. Higher the Sample size, higher the SE.

SE α 1 / (n) Sample size
Where, SE – is the Standard error.

Significance

A distribution of sample that has a smaller SE is a **"Better Estimator of Population Mean"** than a distribution of sample that has a larger SE.

CORRELATION / CORRELATION ANALYSIS

Based on number of variables, there are 3 types of statistical analysis. Viz.

01. Univariate analysis
02. Bivariate analysis
03. Multivariate analysis

Univariate Analysis – The statistical analysis that has only 1 variable, called as Univariate analysis.

e.g. Mean, Mode.

Bivariate Analysis – Those set of analyses which have 2 variables are called as Bivariate Analysis.

e.g. Correlation, Regression analysis.

Multivariate Analysis – Those set of Analysis which have more than 2 variables, are called as Multivariate analysis.

e.g. Multiple correlation analysis, Multiple regression analysis.

CORRELATION

Definition – Correlation is the method of investigating the relationship between the 2 variables. Both of which are quantitative in nature.

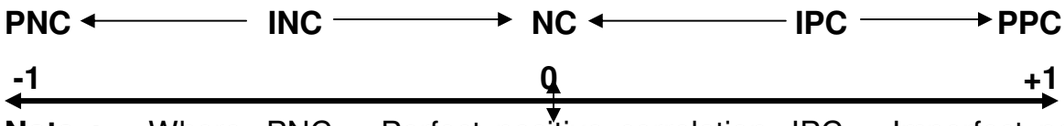
Correlation analysis attempts to determine the degree of two variables.

e.g. Increase of advertisement and increase of sales.

Increase in family income decrease in infant mortality rate.

TYPES

There are five types of correlation.



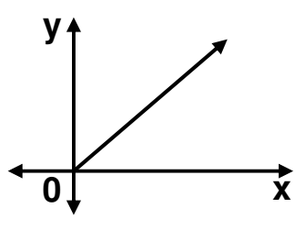
Note : – Where, PNC – Perfect positive correlation. IPC – Imperfect positive correlation, PNC – Perfect Negative correlation. INC – Imperfect Negative correlation, NC = No correlation.

PERFECT POSITIVE CORRELATION

If the values of 2 variables vary in same direction and same proportion, then it is called as Perfect Positive Correlation. **(PPF)**

Here value of r will be +1.

e.g. Age and expenses.

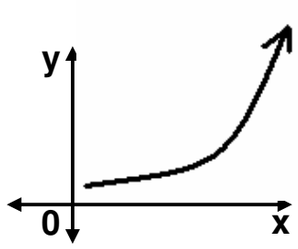


IMPERFECT POSITIVE CORRELATION

If the values of 2 varieties vary in same direction but not in same proportion, then it is called as Imperfect Positive Correlation. (IPC)

Here, value of r will be > +1.

e.g. Income according to the ordinates.

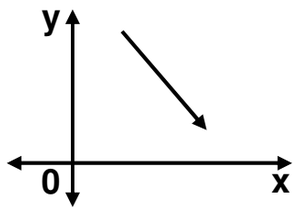


PERFECT NEGATIVE CORRELATION

If the values of 2 variables vary in opposite direction and not in the same proportion, then it is called as Perfect Negative Correlation. (PNC)

Here value of r will be – 1.

e.g. Family income and infant mortality rate.

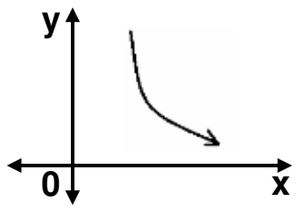


IMPERFECT NEGATIVE CORRELATION

If the value of 2 variables varies in opposite direction but not in same proportion, then it is called as Imperfect Negative correlation.

Here, value of r will be in between - 1 & 0.

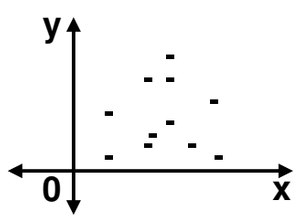
e.g. Number of cigarettes and life span.



NO CORRELATION

If there is no relationship between 2 variables i.e. if the values of 2 variables do not vary either in the same direction or in proportion, then it is called as No Correlation.

e.g. Height of the students and marks scored in exams.



METHODS OF CALCULATION

- Dot / Scattered Diagram.
- Karl Pearson's Coefficient of Correlation.
- Rank Correlation.

KARL PEARSON'S COEFFICIENT OF CORRELATION

It is a mathematical measure of correlation between 2 variables.

It is denoted by the symbol **– r.**

$$r = \frac{\text{Co-variance of x \& y}}{\text{Standard Deviation of x \& y}}$$

$$r = \frac{N(\varepsilon xy) - (\varepsilon x)(\varepsilon y)}{\sqrt{[N(\varepsilon x^2) - (\varepsilon x)^2] \times [N(\varepsilon y^2) - (\varepsilon y)^2]}}$$

Direct Method

Where, r – Coefficient of correlation, N – Number of Variable, x & y – 2 variables, ε – Summation.

$$r = \frac{\text{Co-variance of x \& y}}{\text{Standard Deviation of x \& y}}$$

$$r = \frac{N(\varepsilon uv) - (\varepsilon u)(\varepsilon v)}{\sqrt{[N(\varepsilon u^2) - (\varepsilon u)^2] \times [N(\varepsilon v^2) - (\varepsilon v)^2]}}$$

Indirect Method

Where, r – Coefficient of correlation, N – Number of Variable, x & y – 2 variables, ε – Summation, u & v – Deviated values of x & y respectively. (Where u = x – A & v = y – A. Where A is the assumed value.)

e.g. Following are the height and weight of 10 students. Find the nature of correlation between height and weight.

Age	Weight	Age	Weight
62	50	72	65
78	63	58	50
65	54	70	60
66	61	63	55
60	54	72	65

Answer :

FORMULA FOR DIRECT METHOD

$$r = \frac{\text{Co-variance of x \& y}}{\text{Standard Deviation of x \& y}}$$

$$r = \frac{N(\varepsilon xy) - (\varepsilon x)(\varepsilon y)}{\sqrt{[N(\varepsilon x^2) - (\varepsilon x)^2] \times [N(\varepsilon y^2) - (\varepsilon y)^2]}}$$

Direct Method

Where, r – Coefficient of correlation, N – Number of Variable, x & y – 2 variables, ε – Summation.

Calculate the necessary values in formula –

Age (x)	Weight (y)	xy	x^2	y^2
62	50	3100	3844	2500
72	65	4680	5184	4225
78	63	4914	6084	3969
58	50	2900	3364	2500
65	54	3510	4225	2916
70	60	4200	4900	3600
66	61	4026	4356	3721
63	55	3465	3969	3025
60	54	3240	3600	2915
72	65	4680	5184	4225
ε x = 666	**ε y = 577**	**ε xy = 38715**	**ε x^2 = 44710**	**ε y^2 = 33597**

$(\varepsilon x)^2$ = 443556. **$(\varepsilon y)^2$ =** 332929.

$$r = \frac{10 \times 38715 - 660 \times 577.}{\sqrt{[10\,(44710) - (443556)] \times 10\,(33597) - (332929)]}}$$

$$r = \frac{387150 - 384282.}{\sqrt{[447100 - 443556] \times [335970 - 332929]}}$$

$$r = \frac{2868}{\sqrt{3544 \times 3041}}$$

$$r = \frac{2868}{\sqrt{10777304}}$$

$$r = \frac{2868}{3282.88}$$

r = 0.87.

The correlation in the above given example is – Imperfect Positive Correlation. i.e. There is imperfect positive correlation in Height and Weight in given example.

FORMULA FOR DIRECT METHOD

$$r = \frac{\text{Co-variance of x \& y}}{\text{Standard Deviation of x \& y}}$$

Indirect Method

$$r = \frac{N\,(\varepsilon uv) - (\varepsilon u)\,(\varepsilon v)}{\sqrt{[N\,(\varepsilon u^2) - (\varepsilon u)^2] \times [N\,(\varepsilon v^2) - (\varepsilon v)^2]}}$$

Where, r – Coefficient of correlation, N – Number of Variable, x & y – 2 variables, ε – Summation, u & v – Deviated values of x & y respectively. (Where u = x – A & v = y – A. Where A is the assumed value.)

Calculate the necessary values in formula –

Assumed values of A for u – is 70 & for v – is 60.

x– A	y– A	uv	u^2	v^2
62 – 70 = – 8	50 – 60 = – 10	80	64	100
72 – 70 = 2	65 – 60 = 5	10	4	25
78 – 70 = 8	63 – 60 = 3	24	64	9
58 – 70 = – 12	50 – 60 = – 10	120	144	100
65 – 70 = – 5	54 – 60 = – 6	30	25	36
70 – 70 = 0	60 – 60 = 0	0	0	0
66 – 70 = – 4	61 – 60 = 1	– 4	16	1
63 – 70 = – 7	55 – 60 = – 5	35	49	25
60 – 70 = – 10	54 – 60 = – 6	60	100	36
72 – 70 = 2	65 – 60 = 5	10	4	25
ε u = – 34	**ε v = – 23**	**ε uv = 365**	**ε u^2 = 470**	**ε v^2 = 357**

$(\varepsilon u)^2$ = 1156. **$(\varepsilon v)^2$ =** 529.

$$r = \frac{10 \times 365 - (-34) \times (-23)}{\sqrt{[10\,(470) - (-34)^2] \times 10\,(357) - (-23)^2]}}$$

$$r = \frac{3650 - 782}{\sqrt{[4700 - 1156] \times [3570 - 529]}}$$

$$r = \frac{2868}{\sqrt{3544 \times 3041}}$$

$$r = \frac{2868}{\sqrt{10777304}}$$

$$r = \frac{2868}{3282.88}$$

r = 0.88.

The correlation in the above given example is – <u>IMPERFECT POSITIVE CORRELATION.</u> i.e. There is imperfect positive correlation in Height and Weight in given example.

REGRESSION

It is a bivariate analysis. The word meaning of regression is **"Stepping back or returning to average value."**

The term regression was first introduced in 1877 by a famous British Biometrician **"Sir Franscis Galton."** He studied the relationship between the height of 1000 fathers and Sons and concluded that –

01. All tall fathers had tall sons and all short fathers had short sons.

02. The average height of tall Sons was less than their tall fathers and the average height of short sons was more than their short fathers.

The above study revealed that the height of Sons of abnormally tall or short fathers tend to revert back or step back to the average height of the population. A phenomenon which he described as Regression. But, now-a-days regression is used in wider perspective in the field of statistics.

e.g. Budget, Target setting, etc.

SIGNIFICANCE

Concept of regression is used to predict future events either finding out dependant variable based on independent variable or vice-versa.

REGRESSION EQUATION

01. Regression equation of x on y [Calculation of independent variable (x) based on the dependent variable (y)]

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where, x = Independent variable.

$\bar{x}$ = Arithmetic mean of x series.

b_{xy} = Regression co-efficient of x on y.

b_{xy} is calculated by –

$$b_{xy} = \frac{\varepsilon d_x \times d_y}{\varepsilon d^2_y}$$

Where, y is dependent variable,

$\bar{y}$ is Arithmetic mean of y series.

Where, d_x and d_y are the – deviated values of x and y from its respective arithmetic means. ε = is summation.

02. Regression equation of y on x [Calculation of dependent variable (y) based on independent variable (x)]

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Where,

y – Dependent variable. $\bar{y}$ – is Arithmetic mean.

b_{yx} – Regression co-efficient of y on x.

x – is independent variable. $\bar{x}$ – is Arithmetic mean of x series.

Where, $$b_{yx} = \frac{\varepsilon (d_x \times d_y)}{\varepsilon d^2_x}$$

Where, d_x and d_y = deviated values of x and y from its respective Arithmetic mean. ε = summation.

CALCULATION OF CO-RELATION CO-EFFICIENT USING REGRESSION EQUATION

$$r = \sqrt{b_{xy} \times b_{yx}}$$

r = Co-relation co-efficient.

b_{xy} = Regression co-efficient of x on y.

b_{yx} = Regression co-efficient of y on x.

e.g. Following are the age and systolic blood pressure of 5 patients. Calculate the systolic blood pressure when his age is 45 years. Calculate the age when his systolic blood pressure is 180 mm of Hg. Also calculate co-relation of x and y.

Age	**Systolic blood pressure in mm of Hg**
(x)	**(y)**
40	130
50	150
30	120
20	110
60	160

Answer :

Age	**SBP**	**Mean**	**Mean**	d_x	d^2_x	d_y	d^2_y	$d_x d_y$	b_{xy}	b_{yx}
x	Y	$\bar{x}$	$\bar{y}$	$(x - \bar{x})$		$(y - \bar{y})$				
40	130	40+	130+	0	0	– 4	16	0		
50	150	50+	150+	10	100	+ 16	256	160		
30	120	30+	120+	– 10	100	– 14	196	140		
20	110	20+	110+	– 20	400	– 24	576	480		
60	160	60/5	160/5	20	400	26	676	520		
		40	134	0	1000	0	1720	1300	0.76	1.3

01. <u>Regression equation of y on x</u> [Calculation of dependent variable (y) based on independent variable (x)]

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

Where,

y – Dependent variable = Systolic B.P. $\bar{y}$ – is Arithmetic mean of y series

b_{yx} – Regression co-efficient of y on x.

x – is independent variable = age 45 years. $\bar{x}$ – is Arithmetic mean of x series.

Where, $b_{yx} = \frac{\varepsilon (d_x \times d_y)}{\varepsilon d^2_x}$

Where, d_x and d_y = deviated values of x and y from its respective Arithmetic mean. ε = summation.

$b_{yx} = (1300) / 1000$

$b_{yx} = 1.3$

$y - 134 = 1.3 \times (45 - 40)$

$y - 134 = 1.3 \times 5$

$y = 6.5 + 134$

$y = 140.5$ mm of Hg.

The systolic blood pressure when his age is 45 years will be 140.5 mm of Hg.

02. <u>Regression equation of x on y</u> [Calculation of independent variable (x) based on the dependent variable (y)]

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

Where, x = Independent variable = Age.

$\bar{x}$ = Arithmetic mean of x series.

b_{xy} = Regression co-efficient of x on y.

b_{xy} is calculated by –

$$b_{xy} = \frac{\varepsilon d_x d_y}{\varepsilon d^2_y}$$

Where, y is dependent variable = systolic blood pressure = 180 mm of Hg.

$\bar{y}$ is Arithmetic mean of y series.

Where, d_x and d_y are the – deviated values of x and y from its respective arithmetic means. ε = is summation.

$b_{xy} = 1300 / 1720.$

$b_{xy} = 0.76.$

$x - 40 = 0.76 (180 - 134)$

$x - 40 = 0.76 \times 46$

$x - 40 = 34.96.$

$x = 34.96 + 40.$

x = 74.96.

The systolic blood pressure will be 180 mm of Hg when his age will be approximately 75 years.

CALCULATION OF CO-RELATION CO-EFFICIENT

USING REGRESSION EQUATION

$$r = \sqrt{b_{xy} \times b_{yx}}$$

r = Co-relation co-efficient.

b_{xy} = Regression co-efficient of x on y.

b_{yx} = Regression co-efficient of y on x.

$$r = \sqrt{0.76 \times 1.3}$$

$$r = \sqrt{0.988.}$$

r = 0.993.

The co-relation co-efficient of x and y is type of imperfect positive or near perfect positive co-relation.

TESTS OF SIGNIFICANCE

It enables us to prove or disprove the hypothesis. i.e. Whether it is significant or non-significant and to what extent it is significant.

Definition

It is a measure or tool to prove or disprove the hypothesis.

Hypothesis

It is a tentative conclusions / presumptions which are drawn by the researcher or investigator.

It is of 2 types. Viz. –

- Null hypothesis.
- Research / Alternate hypothesis.

NULL HYPOTHESIS – it is a hypothesis of no effect and formulated with the aim of being rejection. This part takes a great role in implication of any rules and regulations in public or population.

RESEARCH HYPOTHESIS – It is a hypothesis of effect and formulated with the aim of being acceptance.

Test of significance is 2 folded.

- Comparing within the groups.
- Comparing between the groups.

Comparing within the groups – Comparing the results before and after the treatment of same sample.

Comparing between the groups – Comparing the results between the 2 or more groups.

SIX STEPS FOR ALL THE TESTS OF SIGNIFICANCE

01. Formulate the hypothesis. (i.e. both the Null and Research hypothesis)

02. Selection of appropriate type of tests of significance.

- **'t' test** – Calculation in 1 or 2 groups if the number of sample is less than 30.
- **'z' test** – Calculation in 1 or 2 groups if the number of sample is more than 30.
- **'f' test** – Calculation in more than 2 groups and irrespective of sample size.
- **'x^2' test** – To Compare observed values with expected values.

03. Selection of the level of significance.

Decimal	Significance Level	Confidence Level	Remarks
0.1	10%	90%	10 in 100
0.05	5%	95%	5 in 100
0.02	2%	98%	2 in 100
0.01	1%	99%	1 in 100
0.001	0.1%	99.9%	0.1 in 100
0.0001	0.01%	99.99%	0.01 in 100

04. Calculation of sample mean, standard deviation, standard error and any of the selected test of significance i.e. t / f / z / x^2 test.

05. Comparing the observed values with the table value of selected test of significance.

06. Drawing the conclusion based on the above steps.

't' TEST / STUDENT 't' TEST / STUDENT TEST

Among all the test of significance the most common is z test because of larger sample. It is based on standard distribution / normal distribution / Gaussian distribution / Naval distribution. But, when the sample size are less or small (i.e. less than 30) it does not follow normal distribution. **Therefore,** there was a need of a test of significance for smaller samples.

The early work / initial work was done by **W. S. Gossett in Ierland**, who was working in a beverages company. The company did not allow its employ to publish any research article. So he published this test in the pen name of student test.

Therefore, this test became famous by the name of **student test** / **student 't' test** / **'t' test.**

APPLICATIONS

- The samples are randomly selected.
- It should be a quantitative data.
- Variable should be normally distributed. (Symmetrical distribution)
- The sample size should be less than 30.
- When the sample size gets larger than (i.e. more than 30) the t distribution is approximately equal to normal distribution.

TEST OF SIGNIFICANCE

Mainly there are 2 types of t test.

- Unpaired 't' test
- Paired t 't' test

Unpaired 't' test

It is adopted when we want to compare the results between 2 different groups.

Paired 't' test

It is used when we want test of significance of a same sample in different occasions and time like before and after the intervention readings of the same sample. (i.e. within the same group but at different occasions)

UNPAIRED 't' TEST

Calculations

t = Difference in mean of 2 groups / S. E. of 2 groups.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{SE(\bar{x}_1 - \bar{x}_2)}$$

Where, S. E. $(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{(n_1 - 1)\,SD_1^2 + (n_2 - 1)\,SD_2^2}{n_1 + n_2 - 2} \times \frac{1}{n_1} + \frac{1}{n_2}}$

Where, t – is unpaired t value, $\bar{x}_1$ $\bar{x}_2$ – Arithmetic mean of 1st and 2nd group, n_1 & n_2 – Sample size of 1st and 2nd group, SD_1 & SD_2 – Are the variations / Standard deviations of 1st and 2nd group.

Example : Following are the values of birth weight of high socio-economical group and low socio-economical group. Find whether there is a significant difference between 2 groups.

Given Values	Gr. A (High S-E Status)	Gr. B (Low S-E Status)
Sample size (SS)	$n_1 = 15$	$n_2 = 10$
Arithmetic mean (AM)	$\bar{x}_1 = 2.92$	$\bar{x}_2 = 2.26$
Standard deviation (SD)	$SD_1 = 0.27$	$SD_2 = 0.22$

Step 01 : Postulating Hypothesis.

Null Hypothesis – $H_0 = H_1$. There is no significant difference in low and high socio-economic group interns of birth weight.

Research Hypothesis – $H_0 \neq H_1$. There is a significant difference in low and high socio-economic group interns of birth weight.

Step 02 : Selection of test of significance.

2 groups and less than 30 samples **(i.e. 23)**. So, the **unpaired 't' test** should be applied.

Step 03 : Selection of level of significance.

Formula =

t = Difference of mean of 2 groups / S. E. of 2 groups.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{SE(\bar{x}_1 - \bar{x}_2)}$$

Where, S. E. $(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{(n_1 - 1)\,SD_1^2 + (n_2 - 1)\,SD_2^2}{n_1 + n_2 - 2} \times \frac{1}{n_1} + \frac{1}{n_2}}$

Where, t – is unpaired t value, $\bar{x}_1$ $\bar{x}_2$ – Arithmetic mean of 1st and 2nd group, n_1 & n_2 – Sample size of 1st and 2nd group, SD_1 & SD_2 – Are the variations / Standard deviations of 1st and 2nd group.

$$= \sqrt{\frac{(15-1)\,(0.27)^2 + (10-1)\,(0.22)^2 \times 1/15 + 1/10}{15+10-2}}$$

$$= \sqrt{\frac{(14 \times 0.0729) + (9 \times 0.0484) \times 1/15 + 1/10}{15+10-2}}$$

$$= \sqrt{\frac{1.0206 + 0.4356 \times 0.06 + 0.01}{23}}$$

$$= \sqrt{\frac{1.4562 \times 0.16}{23}}$$

$$= \sqrt{\frac{0.2329}{23}}$$

$$= \sqrt{0.010.}$$

$$= \mathbf{0.1006.}$$

<u>Step 04</u> : Calculate the 't' value.

$$t = \frac{|2.92 - 2.26|}{0.1}$$

$$t = 6.6 / 0.1$$

t = 6.6.

<u>Step 05</u> : Compare with the table values.

<u>Degree of freedom</u>

It is calculated by following method. Viz. –

$$\mathbf{n_1 + n_2 - 2.}$$

The obtained **'t' value is 6.6.** By comparing the obtained value with the table value we can get following values. Viz. –

$\mathbf{t_{23,0.05}}$ **=** 2.07.

$\mathbf{t_{23,0.02}}$ **=** 2.50.

$\mathbf{t_{23,0.01}}$ **=** 2.81.

$\mathbf{t_{23,0.001}}$**=** 3.77

<u>Step 06</u> : Drawing the conclusion on the basis of obtained and tabular values for the corresponding values at different levels of significance.

The obtained **'t' value** is **6.6**, which is more than the tale value at the **0.001 significance level (i.e. 3.77),** which is greater than the table value.

Therefore, we have to accept the research hypothesis, which says that there is a significant difference in birth weight of high and low socio-economical status people.

HOMEWORK

Problem

The following data gives the values of acidic reactions of solution (pH test) Test whether there is a significant difference between 2 groups at significant level of **0.001 level**.

Group A	Group B
7	6.8
7.8	7.4
7.9	7
8	7.2
7.6	7.4
7.4	

Step 01 : Postulating Hypothesis.

Null Hypothesis – $H_0 = H_1$. There is no significant difference in group A and group B acid test at significance level of 0.001.

Research Hypothesis – $H_0 \neq H_1$. There is a significant difference in group A and group B acid test at significance level of 0.001.

Step 02 : Selection of level of significance.

2 groups and less than 30 samples **(i.e. 11)**. So, the **unpaired 't' test** should be applied.

Step 03 : Selection of level of significance.

Calculations

$\bar{x}_1$ Group A	$x_1 - \bar{x}_2$	$(\bar{x}_1 - \bar{x}_2)^2$	Group B	$(x - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
7.0	7.0–7.61 = 0.61	0.3721	**6.8**	6.8–7.16 = - 0.36	0.1296
7.8	7.8–7.61 = 0.19	0.361	**7.4**	7.4–7.16 = 0.24	0.0576
7.9	7.9–7.61 = 0.29	0.0841	**7.0**	7.0–7.16 = - 0.16	0.0256
8.0	8.0–7.61 = 0.39	0.1521	**7.2**	7.2–7.16 = 0.04	0.0016
7.6	7.6–7.61 = - 0.01	0.0001	**7.4**	7.4–7.16 = 0.024	0.0576
7.4	7.4–7.61 = -0.21	0.0441			
ε $\bar{x}_1$ = 45.7		0.6886	**ε $\bar{x}_2$ = 35.8**		0.2720

Mean =

$\bar{x} = \varepsilon x / n$

Where, $\bar{x}$ **–** is the Arithmetic mean, **ε** – is the summation, x – is Individual observation, n – is the total number of observations.

Arithmetic Mean of group A = 45.7 / 6 = 7.61.

Arithmetic Mean of group B = 35.8 / 5 = 7.61.

Standard Deviation =

Formula

S.D. $= \sqrt{\varepsilon (x - \bar{x})^2 / n}$.

Where, S.D. – is the Standard Deviation, ε – is the summation, x – is the sum of all individual observations, $\bar{x}$– is the arithmetic mean of the whole group.

SD_1 Standard Deviation of Group A **= 0.6886 / 6 = 0.33.**

SD_2 Standard Deviation of Group B **= 0.2720 / 5 = 0.23.**

Formula =

t = Difference of mean of 2 groups / S. E. of 2 groups.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{SE(\bar{x}_1 - \bar{x}_2)}$$

Where, S. E. $(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{(n_1 - 1)\,SD_1^2 + (n_2 - 1)\,SD_2^2}{n_1 + n_2 - 2} \times \frac{1}{n_1} + \frac{1}{n_2}}$

Where, t – is unpaired t value, $\bar{x}_1$ $\bar{x}_2$ – Arithmetic mean of 1st and 2nd group, n_1 & n_2 – Sample size of 1st and 2nd group, SD_1 & SD_2 – Are the variations / Standard deviations of 1st and 2nd group.

$$= \sqrt{\frac{(6-1)\,(0.33)^2 + (5-1)\,(0.23)^2}{6+5-2} \times 1/6 + 1/5}$$

$$= \sqrt{\frac{(5 \times 0.1089) + (4 \times 0.0529)}{11-2} \times 1/6 + 1/5}$$

$$= \sqrt{\frac{0.5445 + 0.2116}{9} \times 0.16 + 0.2}$$

$$= \sqrt{\frac{0.7561 \times 0.36}{9}}$$

$$= \sqrt{\frac{0.2721}{9}}$$

$$= \sqrt{0.030.}$$

$$= \mathbf{0.1760.}$$

<u>Step 04</u> : Calculate the 't' value.

$$t = \frac{|7.61 - 7.16|}{0.17}$$

$$t = 0.45 / 0.17$$

t = 2.64.

<u>Step 05</u> : Compare with the table values.

The obtained **'t' is 2.64.** By comparing the obtained value with the table value we can get following values. Viz. –

$$\mathbf{t_{11,0.001}} = 4.78$$

<u>Step 06</u> : Drawing the conclusion on the basis of obtained and tabular values for the corresponding values at different levels of significance.

The obtained **'t' value** is **2.64**, which is more than the tale value at the **0.001 significance level (i.e. 4.78),** which is less than the table value.

Therefore, we have to accept the research hypothesis, which says that there is a significant difference in acidic reaction of both the groups.

So, here the **null hypothesis** is accepted, saying that the there is no significant difference in acidic reactions of group A and group B at the significance level of 0.001.

PAIRED 't' TEST

$$t = \frac{|\bar{x} - \mu|}{SE}$$

Where, 't' – is the paired 't' value, $\bar{x}$ - Arithmatic Mean, μ - Population mean or null hypothesis, SE – is Standard Error.

e.g. Following are the results of systolic blood pressure before and after treatment of a hypotensive drug of 9 individuals. Test their significance.

BT	AT	X (I.E. BT – AT)	$X - \bar{x}$	$(X - \bar{x})^2$
122	120	2	2 – 3 = 1	1
121	118	3	3 – 3 = 0	0
120	115	5	5 – 3 = 2	4
115	110	5	5 – 3 = 2	4
126	122	4	4 – 3 = 1	1
130	130	0	0 – 3 = 3	9
120	116	4	4 – 3 = 1	1
125	124	1	1 – 3 = – 2	4
128	125	3	3 – 3 = 0	0
Summation of $(x - \bar{x})^2$				24

STEP 01.: Formulation of Hypothesis.

Null hypothesis – The drug is not having the hypotensive effect.

Research hypothesis – The drug is having the hypotensive effect.

STEP 02. : Selection of test of significance.

Since the sample size is less than 30 and we have to test the significance within the same sample, we have to select the unpaired 't' test.

STEP 03. : Selection of level of significance.

Since, here the level of the significance is not given we have to take it as 0.05.

Decimal	Significant Level	Confidence level	Remarks.
0.05	5%	95%	5 of 100.

STEP 04. : Calculation of standard error.

S.E. = S.D. $/\sqrt{n}$

Where, S.E. – is Standard Error, S.D. – is Standard deviation.

Calculate 't' value.

$$t = \frac{|\bar{x} - \mu|}{SE}$$

Where, 't' – is the paired 't' value, $\bar{x}$ - Arithmatic Mean, μ - Population mean or null hypothesis, SE – is Standard Error.

t value = $\frac{|3 - 0|}{0.54}$

$= 3 / 0.54.$

t value = 5.55.

<u>STEP 05.</u> : Comparison of obtained t value with table value.

Degree of freedom = n – 1.

$= 9 – 1$.

Degree of freedom = 8.

$t_{8,0.05}$ = 2.31.

$t_{8,0.01}$ = 5.01.

<u>STEP 06.</u> : Conclusion.

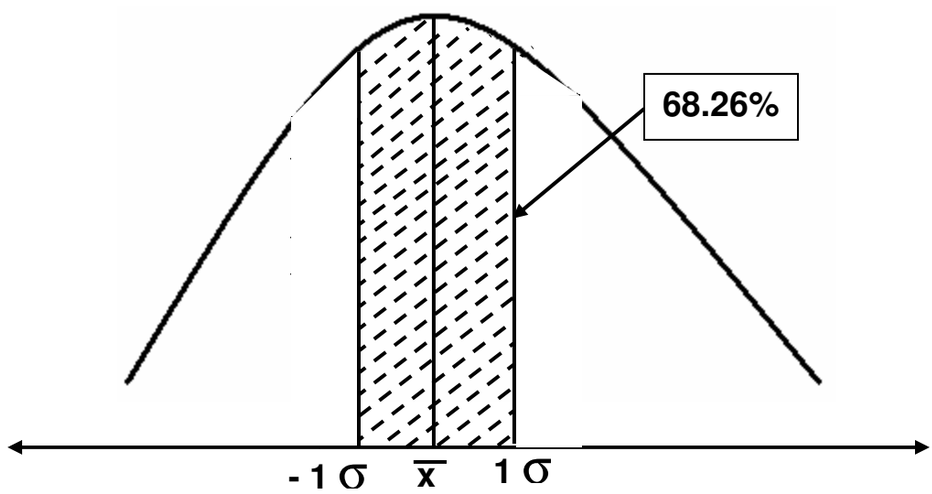
The obtained value is greater than the table value. So, we have to accept the research hypothesis which states that the drug is having hypotensive effect at significant level of 0.01.

NORMAL PROBABILITY CURVE (NPC) / NORMAL DISTRIBUTION / STANDARD DISTRIBUTION / GOUSSION DISTRIBUTION

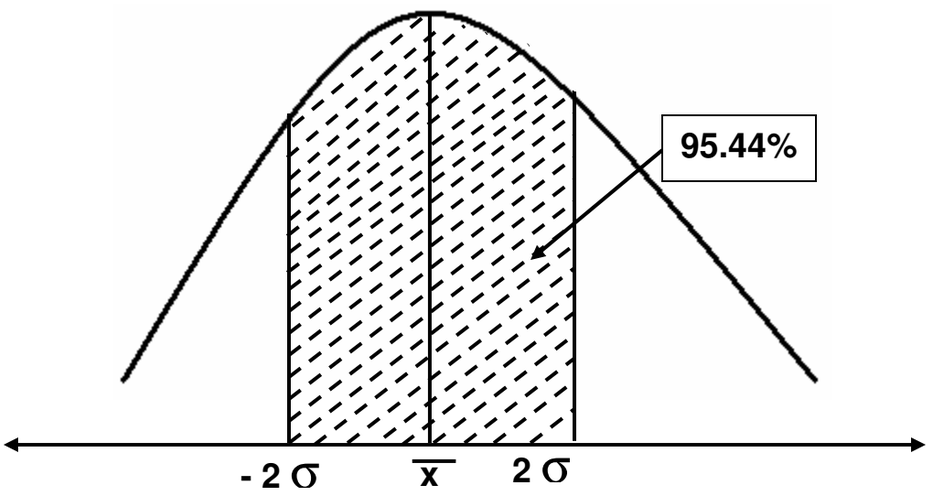
- It is an important continuous probability distribution.
- It is also called as Normal / Standard / Gaussian distribution.
- Between only 2 values assumed by a continuous variable, there exist infinite numbers of variables.
- For such continuous variables the test of significance which is applicable is 'z' test / Normal curve test / Test of significance for larger sample.
- The word probability means – Most likely / High chance.
- The value zero i.e. 0 represents – It will never occur.
- The value one i.e. 1 represents – It is definitely going to occur.
- But this does not occur in the field of biostatistics. In medical field, small number of students and generalized to whole population.

PROPERTIES OF NPC

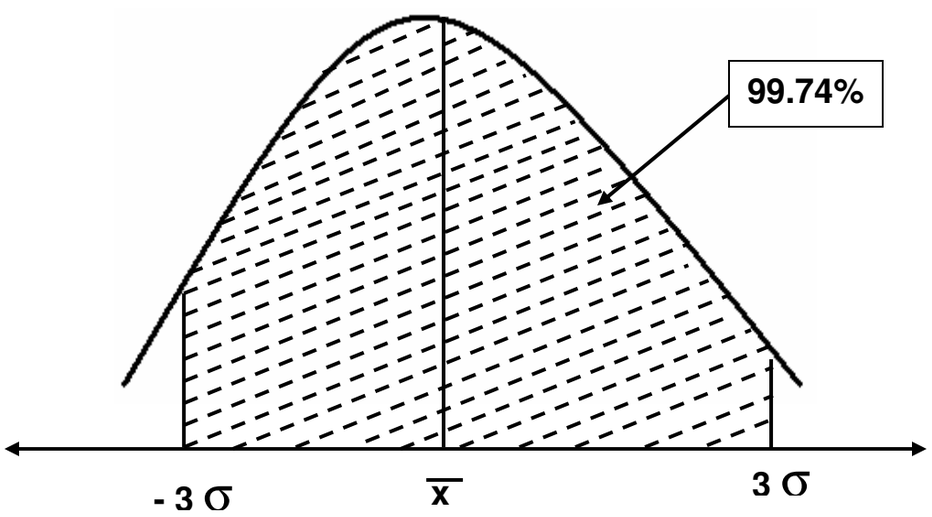
- It is applicable where it is necessary to make inference by taking samples.
- In case of normal distribution – Mean, Median and Mode are same.
- NPC is symmetrically distributed.
- If we draw 2 vertical lines at a distance of +1 or –1 standard deviation from the mean. It will cover 68.26% of the total observations.



- If we extend these vertical lines +2 or –2 standard deviation from the Arithmetic mean, then it will cover 95.44% of the total observation.



- If we further extend these vertical lines to +3 or –3 standard deviation from the Arithmetic mean, then it will cover 99.74% of the total observations.
- It will never be 100%.



'z' TEST / NORMAL CURVE TEST

- It is most widely used test of significance for larger samples. (i.e. Greater than 30.)
- It is based on Normal distribution. (NPC)
- Karl Gouss invented this normal distribution.

SIGNIFICANCE / APPLICATION

- Samples are randomly collected.
- Data should be quantitative in nature.
- Variables are normally distributed.
- Sample size should be more than 30.

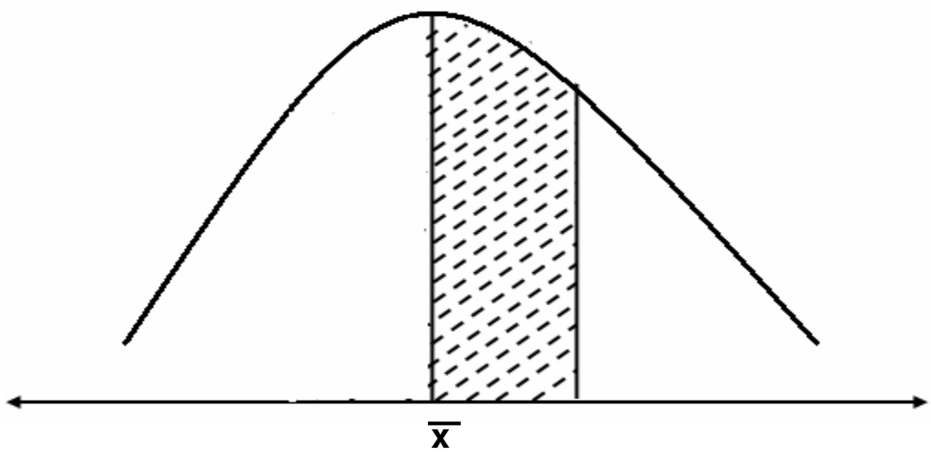
TYPES

There are 2 types of z types.

- One tailed 'z' test.
- Two tailed 'z' test.

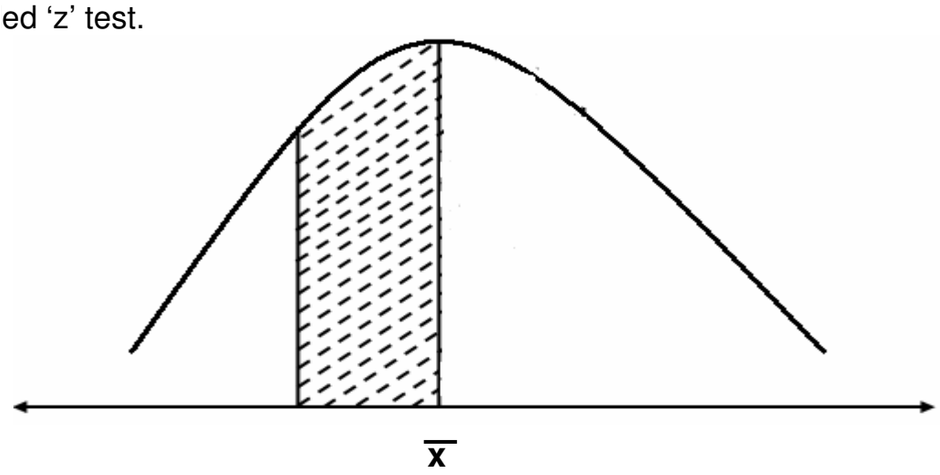
ONE TAILED 'z' TEST

If the distribution is considered only one side, either less than or more than Arithmetic mean, it is called as one tailed 'z' test.



TWO TAILED 'z' TEST

When both sides of the Arithmetic mean are considered then it is called as two tailed 'z' test.



CALCULATION

z value = x – $\overline{x}$ / S. D.

Where, x – Value for which the probability should be calculated.

$\overline{x}$ – Arithmetic mean of the given distribution.

S. D. – Standard deviation.

e.g. A nurse supervisor has found that staff nurses in an average complete a certain task in 10 minutes. If the time required completing a certain task is normally distributed at the standard distribution of 3 minutes. Then calculate –

a) Proportion of nurses completing the task within 4 minutes.

b) Proportion of nurses required less than 5 minutes.

c) Probability that nurses completes the task in between 3 to 6 minutes.

a) For Proportion of nurses completing the task within 4 minutes. (i.e. for <4 minutes)

z value = x – $\overline{x}$ / S. D.

Where, x – Value for which the probability should be calculated.

$\overline{x}$ – Arithmetic mean of the given distribution.

S. D. – Standard deviation.

Here, Arithmetic Mean is 10.

Standard deviation is 3.

Then,

z = 4 – 10 / 3 = – 6 / 3 = – 2.

z = – 2.

'p' value = 0.0228.

In % = 2.28%.

Therefore, about 2.28% of nurses complete the task within 4 minutes.

b) For proportion of nurses required less than 5 minutes. (i.e. for >5 minutes)

z value = x – $\overline{x}$ / S. D.

Where, x – Value for which the probability should be calculated.

$\overline{x}$ – Arithmetic mean of the given distribution.

S. D. – Standard deviation.

Then,

z = 5 – 10 / 3 = – 5 / 3 = – 1.66.

z = – 1.66.

'p' value = 0.0485.

In % = 4.45%

p value for > 5 minute in % = 100 – 4.85 = 95.15%.

Therefore, about 95.15% of nurses complete the task less than 5 minutes.

c) For probability that nurses completes the task in between 3 to 6 minutes.

i) First calculate 'p' value for 3.

z value = x – $\overline{x}$ / S. D.

Where, x – Value for which the probability should be calculated.

$\overline{x}$ – Arithmetic mean of the given distribution.

S. D. – Standard deviation.

Here, x = 3.

Then,

z = 3 – 10 / 3 = – 7 / 3 = – 2.33.

z = – 2.33.

'p' value = 0.0099.

minutes.

i) First calculate 'p' value for 6.

z value = x – $\overline{x}$ / S. D.

Where, x – Value for which the probability should be calculated.

$\overline{x}$ – Arithmetic mean of the given distribution.

S. D. – Standard deviation.

Here, x = 6.

Then,

z = 6 – 10 / 3 = – 4 / 3 = – 1.33.

z = – 1.33.

'p' value = 0.0918.

Therefore, 'p' value in between 3 and 6 minutes =

= 0.0918 – 0.0099 = 0.0819.

In % = 8.19%

Therefore, about 8.19% of nurses probably complete the task in between 3 and 6 minutes.

"f" TEST / ANOVA TEST / ANALYSIS OF VARIANCE

INTRODUCTION

"R. A. Fisher" was a person who invented this test. Therefore, it is called as **"f" test.**

APPLICATION OF "f" TEST

It is used when there are more than 2 groups irrespective of number of samples.

UTILITY OF "f" TEST

It is used to test the significance within the groups and between the groups.

CALCULATIONS

$$\text{f ration} = \frac{\text{Mean square between the groups.}}{\text{Mean square within the groups.}}$$

e.g. The haemoglobin values of 3 groups of children who were fed on 3 different diets are given below. Test whether the mean of these 3 groups differ significantly.

GROUP A	GROUP B	GROUP C
11	8	11
10	11	12
10	9	12
11	8	10
10	8	11
		12

STEP 01.: Formulation of Hypothesis

- Null hypothesis – There is no significant difference between the means of these 3 groups. i.e. $H_0 = A = B = C$.
- Research Hypothesis – These is a significant difference between the means of means of these 3 groups. i.e. $H_1 \neq A \neq B \neq C$.

STEP 02.: Selection of appropriate test of significance.

As there are more than 2 groups, we have to select "f" test.

STEP 03.: Selection of level of significance.

Since, it is not given we will take it as 0.05.

STEP 04.: Calculations.

Sub-step I : Total sum of squares.

a) Sum of all items.

$$\varepsilon X = \varepsilon X_A + \varepsilon X_B + \varepsilon X_C$$

= (11+10+10+11+10) + (8+11+9+8+8) + (11+12+12+10+11+12)

= 52 + 44 + 68.

$\varepsilon x = 164.$

b) Sum of squares of all items.

$\varepsilon x^2 = \varepsilon x^2_A + \varepsilon x^2_B + \varepsilon x^2_C$

$\varepsilon x^2_A = (11)^2 + (10)^2 + (10)^2 + (11)^2 + (10)^2$

= 121 + 100 + 100 + 121 + 100.

= 542.

$\varepsilon x^2_B = (8)^2 + (11)^2 + (9)^2 + (8)^2 + (8)^2$

= 64 + 121 + 81 + 64 + 64.

= 394.

$\varepsilon x^2_C = (11)^2 + (12)^2 + (12)^2 + (10)^2 + (11)^2 + (12)^2$

= 121 + 144 + 144 + 100 + 121 + 144.

= 774.

εx^2 = 542 + 394 + 774.

$\varepsilon x^2 = 1710.$

c) Correction term

Correction term = $(\varepsilon x)^2$ / n.

Where, εx – Total of all items, n – Total number of observations.

Correction term = $(164)^2$ / 16.

= 26896 / 16.

Correction term = 1681.

d) Total sum of squares.

Total sum of squares = Sum of squares of all items – Correction term.

Total sum of squares = 1710 – 1681.

Total sum of squares = 29.

<u>Sub-step II</u> : Total sum of squares between the groups.

a) Squares of total between the groups.

$(\varepsilon x_A)^2 = (52)^2 = 2704.$

$(\varepsilon x_B)^2 = (44)^2 = 1936.$

$(\varepsilon x_C)^2 = (68)^2 = 4624.$

b) Divide by number of observations of each groups.

$$\frac{(\varepsilon x_A)^2}{n_1} = \frac{2704}{5} = 540.8$$

$$\frac{(\varepsilon x_B)^2}{n_2} = \frac{1936}{5} = 387.2$$

$$\frac{(\varepsilon x_C)^2}{n_3} = \frac{4624}{6} = 770.6$$

c) Add the quotients.

$$\frac{(\varepsilon x_A)^2}{n_1} + \frac{(\varepsilon x_B)^2}{n_2} + \frac{(\varepsilon x_C)^2}{n_3}$$

540.8 + 387.2 + 770.6 = 1698.6.

Addition of Quotients = 1698.6.

d) Total sum of squares between the groups.

Total sum of squares between the groups =
Total of quotients – Correction term.

Total sum of squares between the groups = 1698.6 – 1681 = 17.6.

Total sum of squares between the groups = 17.6.

<u>Sub-step III</u> : Total sum of squares within the groups.

Total sum of squares within the groups =
Total sum of squares – Total sum of squares between the groups.

Total sum of squares within the groups = 29 – 17.6 = 11.4.

Total sum of squares within the groups = 11.4.

<u>Sub-step IV</u> : Degree of freedom.

a) Degree of freedom of total sum of square.

Degree of freedom = n – 1.

Degree of freedom = 16 – 1.

Degree of freedom = 15.

b) Degree of freedom of total sum of square between the groups.

Degree of freedom of total sum of square between the groups = K – 1.

Where, K – Number of categories or groups.

Degree of freedom of total sum of square between the groups = 3 – 1.

Degree of freedom of total sum of square between the groups = 2.

c) Degree of freedom of total sum of square within the groups.

Degree of freedom of total sum of square within the groups =
Degree of freedom of total sum of squares – Degree of freedom of total sum of squares between the groups.

Degree of freedom of total sum of square within the groups = 15 – 2.

Degree of freedom of total sum of square within the groups = 13.

<u>Sub-step V</u> : ANOVA table.

$$\text{Mean square} = \frac{\text{Total sum}}{\text{Degree of freedom}}$$

Variation	Total sum	Degree of freedom	Mean square
Between the groups	**17.6**	**2**	17.6 / 2 = **8.8**
Within the groups	**11.4**	**13**	11.4 / 13 = **0.87**

STEP 05.: Comparison of value of **"f" ratio** with **"f" table**.

$f_{13,2,0.05}$ **= 3.80.**

$f_{13,2,0.01}$ **= 6.70.**

Sub-step VI : Calculation of "f" ratio.

$$\text{f ratio} = \frac{\text{Mean square between the groups}}{\text{Mean square within the groups}}$$

f ratio = 8.8 / 0.87 = 10.11.

f ratio = 10.11.

STEP 06.: Conclusion.

Since, the obtained "f" ratio value is more than the table "f" value at significant level of 0.05 and 0.01.

So, we have to accept **RESEARCH HYPOTHESIS** which states that there is a significant difference in the Hb% of the 3 groups who were fed on 3 different diets.

HOMEWORK

The following are the weights of 4 groups. Test whether they differ significantly.

GROUP A	GROUP B	GROUP C	GROUP D
6	8	3	4
4	5	9	5
8	7	6	8
3		5	7

STEP 01.: Formulation of Hypothesis

- Null hypothesis – There is no significant difference between the means of these 4 groups. i.e. $H_0 = A = B = C = D$.
- Research Hypothesis – These is a significant difference between the means of means of these 4 groups. i.e. $H_1 \neq A \neq B \neq C \neq D$.

STEP 02.: Selection of appropriate test of significance.

As there are more than 2 groups, we have to select "f" test.

STEP 03.: Selection of level of significance.

Since, it is not given we will take it as 0.05.

STEP 04.: Calculations.

Sub-step I : Total sum of squares.

a) Sum of all items.

$\varepsilon x = \varepsilon x_A + \varepsilon x_B + \varepsilon x_C + \varepsilon x_D$

$= (6+4+8+3) + (8+5+7) + (3+9+6+5) + (4 + 5 + 8 + 7)$

$= 21 + 20 + 23 + 24.$

$\varepsilon x = 88.$

b) Sum of squares of all items.

$\varepsilon x^2 = \varepsilon x^2_A + \varepsilon x^2_B + \varepsilon x^2_C + \varepsilon x^2_D$

$\varepsilon x^2_A = (6)^2 + (4)^2 + (8)^2 + (3)^2$

$= 36 + 16 + 64 + 9.$

$= 125.$

$\varepsilon x^2_B = (8)^2 + (5)^2 + (7)^2$

$= 64 + 25 + 49.$

$= 138.$

$\varepsilon x^2_C = (3)^2 + (9)^2 + (6)^2 + (5)^2$

$= 9 + 81 + 36 + 25.$

$= 151.$

$\varepsilon x^2_D = (4)^2 + (5)^2 + (8)^2 + (7)^2$

$= 16 + 25 + 64 + 49.$

$= 154.$

$\varepsilon x^2 = 125 + 138 + 151 + 154.$

$\varepsilon x^2 = 568.$

c) Correction term

Correction term = $(\varepsilon x)^2$ / n.

Where, εx – Total of all items, n – Total number of observations.

Correction term = $(88)^2 / 15.$

$= 7744 / 15.$

Correction term = 516.26.

d) Total sum of squares.

Total sum of squares = Sum of squares of all items – Correction term.

Total sum of squares = 568 – 516.26 = 51.74.

Total sum of squares = 51.74.

<u>Sub-step II</u> : Total sum of squares between the groups.

a) Squares of total between the groups.

$(\varepsilon x_A)^2 = (21)^2 = 441.$

$(\varepsilon x_B)^2 = (20)^2 = 400.$

$(\varepsilon x_C)^2 = (23)^2 = 529.$

$(\varepsilon x_D)^2 = (24)^2 = 576.$

b) Divide by number of observations of each groups.

$$\frac{(\varepsilon x_A)^2}{n_1} = \frac{441}{4} = 110.25.$$

$$\frac{(\varepsilon x_B)^2}{n_2} = \frac{400}{3} = 133.33.$$

$$\frac{(\varepsilon x_C)^2}{n_3} = \frac{529}{4} = 132.25.$$

$$\frac{(\varepsilon x_D)^2}{n_2} = \frac{576}{4} = 144.$$

c) Add the quotients.

$$\frac{(\varepsilon x_A)^2}{n_1} + \frac{(\varepsilon x_B)^2}{n_2} + \frac{(\varepsilon x_C)^2}{n_3} + \frac{(\varepsilon x_D)^2}{n_4}$$

$110.25 + 133.33 + 132.25 + 144 = 519.83.$

Addition of Quotients = 519.83.

d) Total sum of squares between the groups.

Total sum of squares between the groups =
Total of quotients – Correction term.

Total sum of squares between the groups = 519.83 – 516.26 = 3.57.

Total sum of squares between the groups = 3.57.

<u>Sub-step III</u> : Total sum of squares within the groups.

Total sum of squares within the groups =
Total sum of squares – Total sum of squares between the groups.

Total sum of squares within the groups = 51.74 – 3.57 = 48.17.

Total sum of squares within the groups = 48.17.

<u>Sub-step IV</u> : Degree of freedom.

a) Degree of freedom of total sum of square.

Degree of freedom = n – 1.

Degree of freedom = 15 – 1 = 14.

Degree of freedom = 14.

b) Degree of freedom of total sum of square between the groups.

Degree of freedom of total sum of square between the groups = K – 1.

Where, K – Number of categories or groups.

Degree of freedom of total sum of square between the groups = 4 – 1 = 3.

Degree of freedom of total sum of square between the groups = 3.

c) Degree of freedom of total sum of square within the groups.

Degree of freedom of total sum of square within the groups = Degree of freedom of total sum of squares – Degree of freedom of total sum of squares between the groups.

Degree of freedom of total sum of square within the groups = 14 – 3 = 11.

Degree of freedom of total sum of square within the groups = 11.

Sub-step V : ANOVA table.

$$\text{Mean square} = \frac{\text{Total sum}}{\text{Degree of freedom}}$$

Variation	Total sum	Degree of freedom	Mean square
Between the groups	**3.57**	**3**	3.57 / 3 = **1.19**
Within the groups	**48.17**	**11**	48.17 / 11 = **4.37**

STEP 05.: Comparison of value of "f" ratio with "f" table.

$f_{11,3,0.05}$ = 3.59.

$f_{11,3,0.01}$ = 6.22.

Sub-step VI : Calculation of "f" ratio.

$$\text{f ratio} = \frac{\text{Mean square between the groups}}{\text{Mean square within the groups}}$$

f ratio = 1.19 / 4.37 = 0.27.

f ratio = 0.27.

STEP 06.: Conclusion.

Since, the obtained "f" ratio value is less than the table "f" value at significant level of 0.05 and 0.01.

So, we have to accept **NULL HYPOTHESIS** which states that there is a significant difference in the weight of the 4 groups.

CHISQUARE TEST ("x^2" TEST)

INTRODUCTION

- The letter **"x"** in Greek represents **"chi"**. As it is **"x^2"** or square of **"x"** it is called as **"Chisquare test."**
- It was first introduced by a famous statistician **"Karl Pierson"** in 1889.
- It is used for more than 2 categories of data. (i.e. Dichotomus data)
  e.g. Boys and Girls, Yes and No, Rural and Urban, etc.
- It is used to check the prevalence among the data.

APPLICATION / UTILITY

It evaluates whether the observed frequency in a sample differ significantly from the expected frequencies. In other words, it is used to test whether a significant difference exists between the observed number of samples and the expected number of responses.

CALCULATIONS

It is the summation of the squared deviations of each observed frequency from its expected frequency divided by corresponding expected frequency.

$$x^2 = \varepsilon \frac{(O - E)^2}{E}$$

Where, x^2 – Chisquare value, O – Observed value, E – Expected value,

ε – Summation.

INTERPRETATION

It is the difference of Observed value and Expected value is zero or less, then there is no significant difference. But, if the difference is more then, there will be statistically significant difference.

e.g. A doctor has a hypothesis that headache is common among males and females during examinations in a sample of 100 students. If he finds 58 girls and 42 boys suffering from headache, does the finding support or contradict his hypothesis?

STEP 01. : Formulation of Hypothesis.

Null hypothesis – There is no difference between the boys and girls suffering from headache . H_0 = B = G.

Research Hypothesis – There is a significant difference between the boys and girls from headache. H_1 = B = G.

STEP 02. : Selection of appropriate test of significance.

As we have to compare the observed and expected value, we have to select x^2 test.

STEP 03. : Selection of level of significance.

Since, it is not mentioned, we will take it as 0.05.

STEP 04. : Calculations.

$$x^2 = \frac{\varepsilon (O - E)^2}{E}$$

Where, x^2 – Chisquare value, O – Observed value, E – Expected value, ε – Summation.

	EXPECTED VALUES	OBSERVED VALUES
Boys	**50**	**42**
Girls	**50**	**58**

$$x^2 = \frac{\varepsilon (O_B - E_B)^2}{E_B} + \frac{\varepsilon (O_G - E_G)^2}{E_G}$$

$$x^2 = \frac{\varepsilon (42 - 50)^2}{50} + \frac{\varepsilon (58 - 50)^2}{50}$$

$$x^2 = \frac{(8)^2}{50} + \frac{(8)^2}{50}$$

$$x^2 = 128 / 50$$

$$x^2 = 2.56.$$

STEP 05.: Comparison of obtained x^2 value with table value.

$D_f = K - 1 = 2 - 1 = 1.$

$x^2_{1, 0.05} = 3.84.$

STEP 06.: Conclusion.

As the obtained x^2 value is less than the table value, we have to accept null hypothesis, which states that, there is no significant difference between the boys and girls suffering from Headache.

Thus, the statistics support the doctor's hypothesis, which is saying that Headache is common among males and females during examinations.

VITAL STATISTICS

INTRODUCTION

- It is an important branch of biostatistics which is necessary for documentary and legal purpose.
- In India, office of registrar general of India, (RGI) was established in the year 1951 for colleting vital statistics and conducting census.
- The registration of birth and death was made compulsory and uniform all over India in 1969.

DEFINITION

The branch of biostatistics which deals with the important events of the life like birth, death, marriage, etc is called as vital statistics.

USES OR SIGNIFICANCE OF THE VITAL STATISTICS

- To describe the community health.
- To diagnose the community illness.
- To find the solutions for social problems.
- To plan or modify health programmes.
- For maintenance of records.

BASIC REPRESENTATION OF VITAL STATISTICS

It is expressed either in terms of rate or ratio.

RATE

It refers to those calculations that involve frequency of occurrence of some events in a specific period.

It is calculated by –

$$\text{Rate} = \frac{a}{a+b} \times k$$

Where, a – is Frequency of the event during specific period of time, a + b – It is the persons who are exposed to risk of events, k – is the constant, it is generally taken as 1000.

RATIO

It is the proportion between 2 or more events.

e.g. Male and Female ratio in a class, Patient and Doctors ratio in a city, Student and Teacher ratio in a college, etc.

All these can be expressed in 3 index. They are Viz. –

- Mortality
- Morbidity
- Fertility

MORTALITY

Death and birth are unique (i.e. it occurs only once). Hence, its recording is easy.

ACDR = Annual Crude Death Rate.

$$\text{ACDR} = \frac{\text{Total number of death during the year}}{\text{Total mid year population}} \times 1000$$

AIMR = Annual Infant Mortality Rate.

$$\text{AIMR} = \frac{\text{Number of death within 1 year of birth}}{\text{Total number of live births during the year}} \times 1000$$

MORBIDITY

It is difficult to record morbidity. Hence, WHO has laid down few guidelines for recording morbidity. They are Viz. –

- Person
- Illness
- Spells of illness
- Duration

FERTILITY

AFR = Annual Fertility Rate.

$$\text{AFR} = \frac{\text{Number of births during the year}}{\text{Number of females in reproductive age}} \times 1000$$